

Award Number: W81XWH-07-1-0242

TITLE: Infrared Spectroscopic Imaging for Prostate Pathology Practice.

PRINCIPAL INVESTIGATOR: Rohit Bhargava, PhD

CONTRACTING ORGANIZATION: U^{\{æ\b\}}\tilde{A}\sim\tilde{a}\emptyset\rightarrow\times\wedge\leadsto b

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAUUääá^áÊ∅QÁÁ

REPORT DATE: March 2010

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT:

* Approved for public release; distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.					
1. REPORT DATE (DD-MM-YYYY) €F-03-2010		2. REPORT TYPE Annual		3. DATES COVERED (From - To) Feb 15, 2009-Feb 14, 2010	
4. TITLE AND SUBTITLE Infrared Spectroscopic Imaging for Prostate Pathology Rtceveg				5a. CONTRACT NUMBER W81XWH-07-1-0242	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Rohit Bhargava Email:rxb@uiuc.edu				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of Illinois 1901 First St. Suite A Champaign, IL 61820				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) US Army medical research and Materiel Command Fort Detrick, MD 21702-5012				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution unlimited					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT The report summarizes progress towards using Fourier transform infrared spectroscopic imaging for prostate pathology in year 3 of a 3 year award from the PCRP. The aim of the work is to enable histopathologic recognition without the use of human input or stains. The major accomplishments in the past year are:					
15. SUBJECT TERMS Spectroscopy, prostate, histopathology, cancer, optimization, optical imaging					
16. SECURITY CLASSIFICATION OF: U			17. LIMITATION OF ABSTRACT UU	18. NUMBER OF PAGES 116	19a. NAME OF RESPONSIBLE PERSON USAMRMC
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U			19b. TELEPHONE NUMBER (include area code)

Á
Á
Á

Table of Contents

	<u>Page</u>
Introduction.....	3
Body.....	4
Key Research Accomplishments.....	13
Reportable Outcomes.....	25
Conclusion.....	28
References.....	28
Appendices.....	29

Introduction

Prostate cancer accounts for one-third of noncutaneous cancers diagnosed in US men,¹ is a leading cause of cancer-related death and is, appropriately, the subject of heightened public awareness and widespread screening. If prostate-specific antigen (PSA)² or digital rectal screens are abnormal,³ a biopsy is considered to detect or rule out cancer. Pathologic status of biopsied tissue forms the definitive diagnosis for prostate cancer and constitutes an important cornerstone of therapy and prognosis.⁴ There is, hence, a need to add useful information to diagnoses and to introduce new technologies that allow efficient analyses of cancer to focus limited healthcare resources. For the reasons outlined above, there is a urgent need for high-throughput, automated and objective pathology tools. Our general hypothesis is that these requirements are satisfied through innovative spectroscopic imaging approaches that are compatible with, and add substantially to, current pathology practice. Hence, the overall aim of this project is to demonstrate the utility of novel Fourier transform infrared (FTIR) spectroscopy-based, computer-aided diagnoses for prostate cancer and develop the required microscopy and software tools to enable its application.

FTIR spectroscopic imaging is a new technique that combines the spatial specificity of optical microscopy and the biochemical content of spectroscopy.⁵ As opposed to thermal infrared imaging, FTIR imaging measures the absorption properties of tissue through a spectrum consisting of (typically) 1024 to 2048 wavelength elements per pixel.⁶ Since mid-IR (2-12 μm wavelength) spectra reflect the molecular composition of the tissue, image contrast arises from differences in endogenous chemical species. As opposed to visible microscopy of stained tissue that requires a human eye to detect changes, numerical computation is required to extract information from IR spectra of unstained tissue. Extracted information, based on a computer algorithm, is inherently objective and automated. Recent work has demonstrated that these determinations are also accurate and reproducible in large patient populations.⁷ Hence, we focused, in the first year of this project, on demonstrating that the laboratory results could be optimized using novel approaches to fast imaging. This is a critical step, since we propose next to analyze 375 radical prostatectomy samples. We have been able to optimize data acquisition parameters and develop a novel algorithm for processing data that enables almost 50-fold faster imaging. Briefly, the idea behind the process is illustrated in Figure 1. In this performance period, we sought to use acquired data to establish the use of IR imaging for validating cancer diagnosis (task 2), develop a calibration and prediction model for grading and perform extensive validation (task 2). Finally, we sought to develop a mathematical framework to relate disparate pieces of information to outcome (task 3).

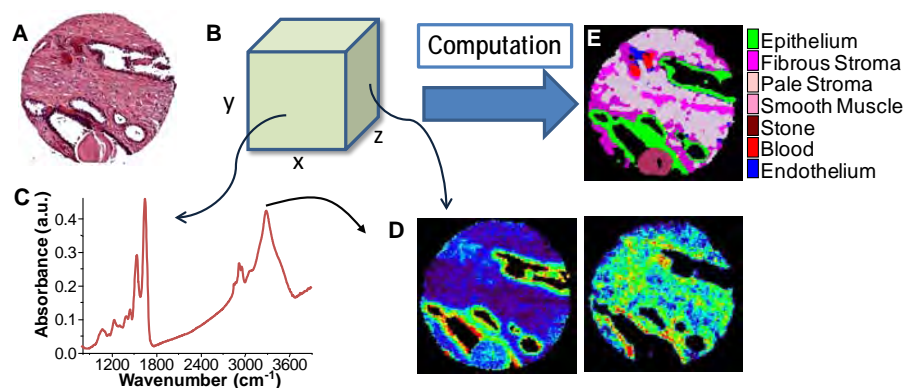


Figure 1. (A) Conventional imaging in pathology requires dyes and a human to recognize cells. In chemical imaging data cubes (B), both a spectrum at any pixel (C) and the spatial distribution of any spectral feature can be seen. e.g. in (D) nucleic acids (left, at $\sim 1080\text{ cm}^{-1}$), and collagen specific (right, at $\sim 1245\text{ cm}^{-1}$) Computational tools can then convert chemical imaging data to knowledge used in pathology (E).

Body

Specific activities and tasks as per statement of work during this performance period are described below. Details of performance for the past years periods are given in the past annual reports which is attached for quick reference of the reviewers. :

Task 1. Perform infrared spectroscopic imaging on prostate biopsy specimens

All activities for this task were completed in year 1 and 2

Task 2. Analyze spectroscopic imaging data for biochemical markers of tumor and develop numerical algorithms for grading cancer

Goal: Develop algorithm for malignancy recognition. Models will be constructed and optimized using Genetic Algorithms operating on identified metrics. Models will be tested and validated using ROC curves with pathologist marking as the ground truth. A protocol for segmenting benign from atypical condition will be available. (Months 11-18) Three specific aims from the statement of work (SOW) are:

- Identify samples to be imaged (Months 1-3) by examining stained slides
- Obtain unstained samples to be imaged and define regions for calibration and validation (Months 4-7)
- Perform histologic identification on prostate samples and validate
- Reduce spectral metrics to those useful in identifying atypia (Months 8-12)
- Develop protocols and validate distinction between benign-appearing and atypical tissue (Months 12-18)
- Develop calibration for predicting cancer grade (Months 18-22)
- Develop protocols and validate Gleason grading of tumor (Months 18-27)

Activities: Task 2a-2d were accomplished in years 1 and 2.

TASK 2E: DEVELOP PROTOCOLS AND VALIDATE DISTINCTION BETWEEN BENIGN-APPEARING AND ATYPICAL TISSUE

We were able to accomplish task 2e entirely and a manuscript has been submitted (under review). An invention disclosure was filed with the office and technology management, who then decided to file a preliminary patent on the work.

We develop a new fully-automated method to classify cancer versus non-cancer prostate tissue samples. The classification algorithm uses morphological features – geometric properties of epithelial cells/nuclei and lumens – that are quantified based on H&E stained images as well as FT-IR images of the samples. By restricting the features used to geometric measures, we sought to mimic the pattern recognition process employed by human experts, and achieve a robust classification procedure that can produce consistently high accuracy across independent data sets. We systematically evaluate the performance of the new method through cross-validation, and examine its robustness across data sets. We also summarize the specific morphological features that prove to be most informative in classification.

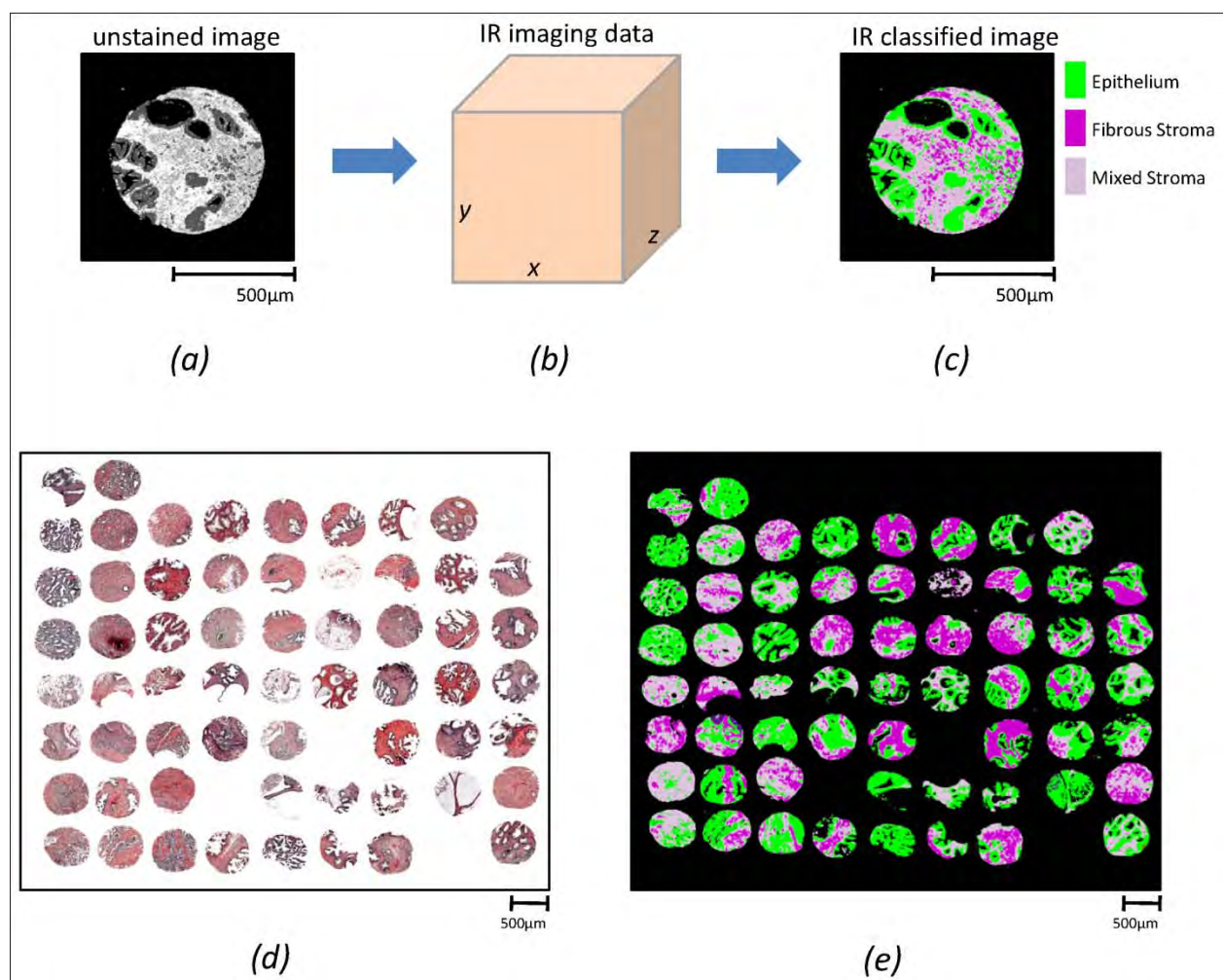


Figure 2. IR imaging data and its use in histologic classification.

(Upper row) IR imaging data (b) is acquired for an unstained tissue section (a). The data is then classified into cell types and a classified image (c) is obtained. The colors indicate cell types in a histologic model of prostate tissue. This method is robust and applied to hundreds of tissue samples using the tissue microarray (TMA) format. (Lower row) H&E (d) and IR classified (e) images of a part of the TMAs used.

Methods: Several new methods were developed to accomplish the task.

We begin with a description of the computational pipeline. As noted above, a key aspect of our approach is the use of FT-IR imaging data on a serial section that is H&E-stained to enhance the segmentation of nuclei and lumens. The first two components of the pipeline (§1-2) are geared to this functionality, while the next three components (§3-5) exploit the segmented features obtained from image data to classify the tissue sample (Figure 3).

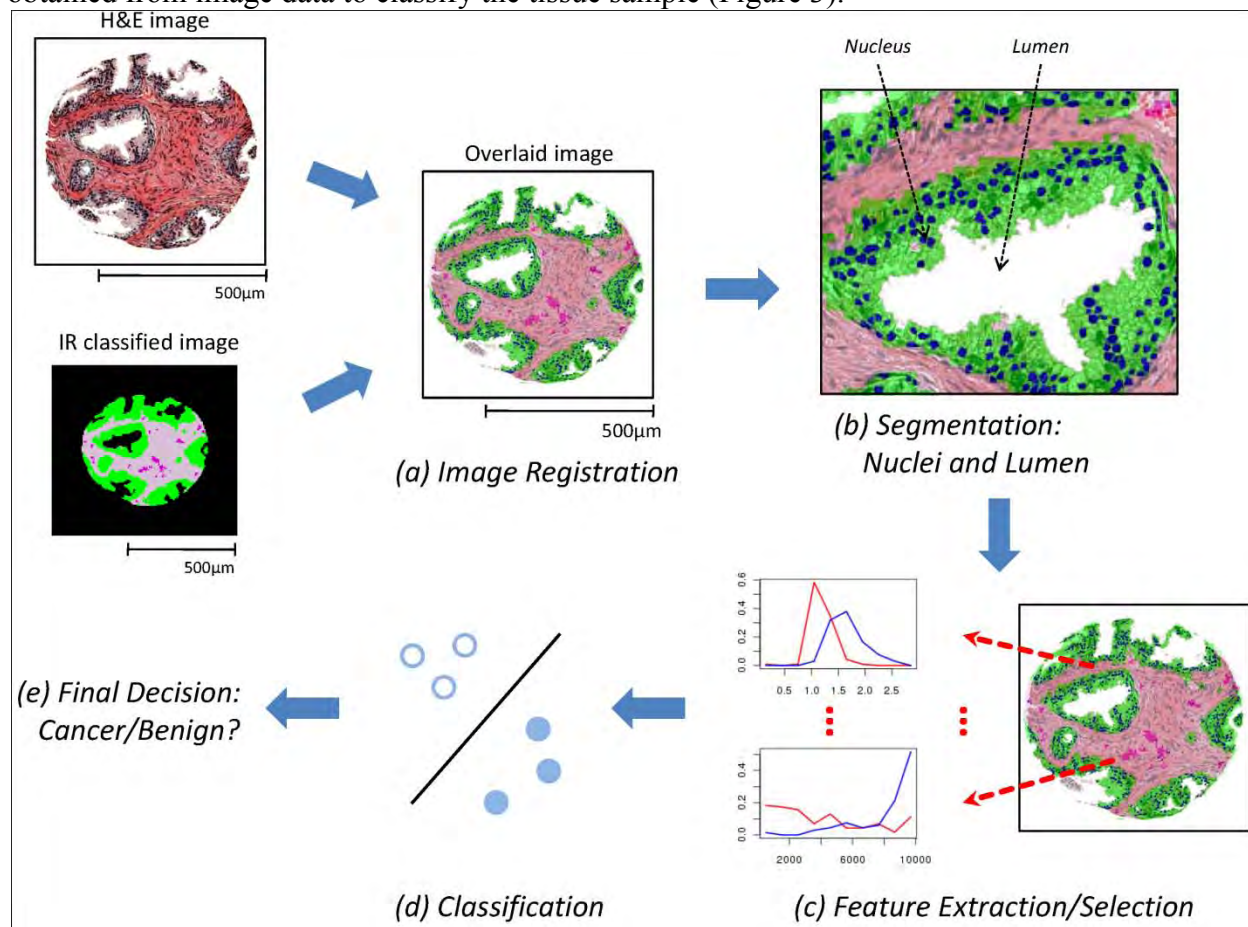


Figure 3. Overview of the approach.

(a, b) FTIR spectroscopic imaging data-based cell-type classification (IR classified image), is overlaid with H&E stained image (a), leading to segmentation of nuclei and lumens in a tissue sample (b). (c,d,e) Features are extracted and selected (c), and used by the classifier (d) to predict (e) whether the sample is cancerous or benign.

1. Image Registration

Given two images, the image registration problem can be defined as finding the optimal spatial and intensity transformation of one image to the other. Here, two images are H&E stained and “IR classified” images which were acquired from adjacent tissue samples. The IR classified image represents the FT-IR imaging data, processed as indicated in Figure 2, to classify each pixel as a particular cell type. Although the two samples were physically in the same intact tissue and are structurally similar, the two images have different properties (total image and pixel sizes, contrast mechanisms and data values). Hence, features to spatially register the images are not trivial. The H&E image provides detailed morphological information that could ordinarily be used for registration, but the IR image lacks such information. On the other hand, the IR image specifies the exact areas corresponding to each cell type, but the difficulty in precisely extracting such regions from the H&E image hinders us from using cell-type information for registration. The only obvious features are macroscopic sample shape and empty space (lumens) inside the samples. To utilize these two features and to avoid problems due to differences in the two imaging techniques, both images are first converted into binary images. Due to the binarization, the intensity transformation is not necessary. As a spatial transformation, we use an affine transformation (f) where a coordinate (x_1, y_1) is transformed to the (x_2, y_2) coordinate after translations (t_x, t_y) , rotation by θ , and scaling by factor s .

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} t_x \\ t_y \end{bmatrix} + s \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$

Accordingly, we find the optimal parameters of the affine transformation that minimizes the absolute intensity difference between two images ($I_{reference}$ and I_{target}). In other words, image registration amounts to finding the optimal parameter values $(t_x^*, t_y^*, \theta^*, s^*) = \arg \min_{t_x, t_y, \theta, s} |I_{reference} - f(I_{target}; t_x, t_y, \theta, s)|$. The downhill simplex method is applied to solve the above equation. An example of this registration process is shown in Figure 4.

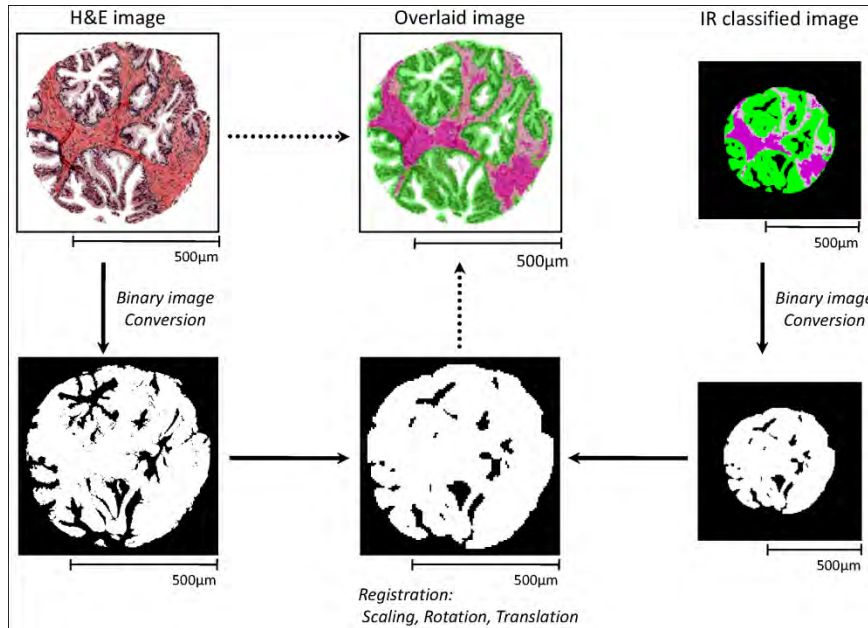


Figure 4. Image Registration.

H&E stained images and IR classified images are first converted into binary images. The IR classified image is overlaid with the H&E stained image by affine transformation, with the optimal matching being found by minimizing the absolute intensity difference between two images. After registration, original annotations (color and/or cell-type information) of each image are restored.

2. Identification of epithelial cells and their morphologic features

While a number of factors are known to be transformed in cancerous tissues, epithelial morphology is utilized as the clinical gold standard. Hence, we focus here on cellular and nuclear morphology of epithelial nuclei and lumens. These structures are different in normal and cancerous tissues, but are not widely used in automated analysis due to a few reasons. First, as described above, simple detection of epithelium from H&E images is difficult. Second, detection of epithelial nuclei may be confounded by a stromal response that is not uniform for all grades and types of cancers. We focused first on addressing these two challenges that hinder automatically parsing morphologic features such as the size and number of epithelial nuclei and lumens, distance from nuclei to lumens, geometry of the nuclei and lumens, and others (§3). In order to use these properties, the first step is to detect nuclei and lumens correctly and we sought to develop a robust strategy for the same.

2.1. Lumen Detection

In H&E stained images, lumens are recognized to be empty white spaces surrounded by epithelial cells. In normal tissues, lumens are larger in diameter and can have a variety of shapes. In cancerous tissues, lumens are progressively smaller with increasing grade and generally have less distorted elliptical or circular shapes. Our strategy to detect lumens was to find empty areas that are located next to the areas rich in epithelium. White spots inside the sample can be found from the H&E image, and the pixels corresponding to epithelial cells can be mapped on the H&E image from the IR classified image through image registration. We note that while lumens are ideally completely surrounded by epithelial cells (called complete lumens), some samples have lumens (called incomplete lumens) that violate this criterion because only a part of lumen is present in the sample. To identify these incomplete lumens, we use heuristic criteria based on the size, shape, presence of epithelial cells and background around the areas, and distance from the center of the tissue. (See Supplementary Materials for details.)

2.2. Nucleus Detection – single epithelial cells

Epithelial nucleus detection by automated analysis is more difficult than lumen detection due to variability in staining and experimental conditions under which the entire set of H&E images were acquired. Differences between normal and cancerous tissues, and among different grades of cancerous tissues, also hamper facile detection. To handle such variations and make the contrast of the images consistent, we perform smoothing and adaptive histogram equalization prior to nuclei identification. Nuclei are relatively dark and can be modeled as small elliptical areas in the stained images. This geometrical model is often confounded as multiple nuclei can be so close as to appear like one large, arbitrary-shaped nucleus. Also, small folds or edge staining around lumens can make the darker shaded regions difficult to analyze. Here, we exploit the information provided by the IR classified image to limit ourselves to epithelial cells, and use a thresholding heuristic on a color space-transformed image to identify nuclei with high accuracy. Epithelial pixels that are identified on the H&E images using the IR overlay provide pixels of dominated by

one of two colors: blue or pink, which arise from the nuclear and cytoplasmic component respectively. For nuclei restricted to epithelial cells in this manner, a set of general observations were made that led us to convert the stained image to a new color space “RG-B” ($|R + G - B|$). (R, G, and B represent the intensity of Red, Green, and Blue channels, respectively.) This transformation, followed by suitable thresholding, was able to successfully characterize the areas where nuclei are present. The threshold values are adaptively determined for Red and Green channels due to the variations in the color intensity. (See Supplementary Materials for details.) Finally, filling holes and gaps within nuclei by a morphological closing operation, the segmentation of each nucleus is accomplished by using a watershed algorithm followed by elimination of false detections. The size, shape, and average intensity are considered to identify and remove artifactual nuclei. Figure 5 details the nucleus detection procedure.

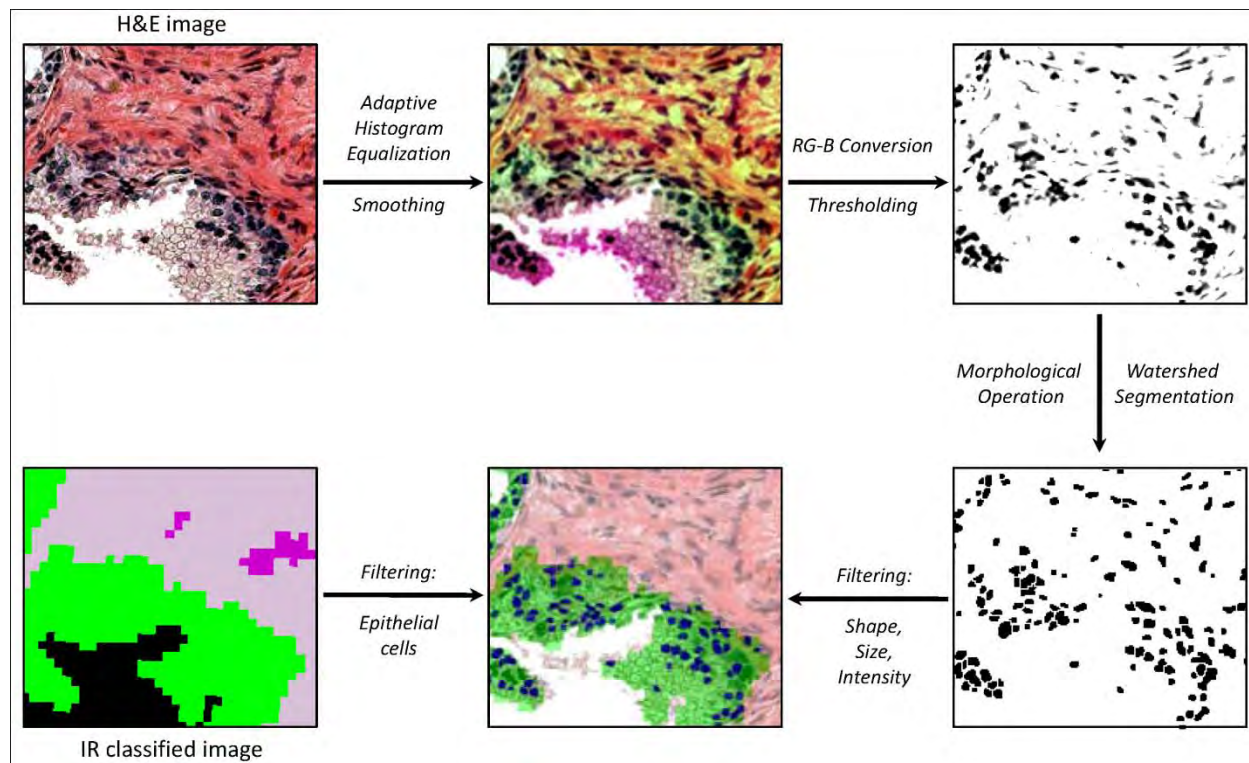


Figure 5. Nucleus Detection.

Smoothing and adaptive histogram equalization are performed to alleviate variability in H&E stained image and to obtain better contrast. “RG – B” conversion followed by thresholding characterizes the areas where nuclei exist. Morphological closing operation is performed to fill holes and gaps within nuclei, and a watershed algorithm segments each individual nuclei. The segmented nuclei are constrained by their shape, size, and average intensity and epithelial cell classification (green pixels) provided by the overlaid IR image.

3. Feature Extraction

As mentioned above, the characteristics of nuclei and lumens change in cancerous tissues. In a normal tissue, epithelial cells are located mostly in thin layers around lumens. In cancerous tissue, these cells generally grow to fill lumens, resulting in a decrease in the size of lumens, with

the shape of lumens becoming more elliptical or circular. The epithelial association with a lumen becomes inconsistent and epithelial foci may adjoin lumens or may also exist without an apparent lumen. Epithelial cells invading the extra-cellular matrix also result in a deviation from the well-formed lumen structure; this is well-recognized as a hallmark of cancer. Due to filling lumen space and invasion into the extra-cellular space, the number density of epithelial cells increases in tissue. The size of individual epithelial cells and their nuclei also tend to increase as malignancy of a tumor increases. Motivated by such recognized morphological differences between normal and cancerous tissues, we chose to use epithelial nuclei and lumens as the basis of the several quantitative features that our classification system works with. (See examples of such features in Figure 6.) It is notable that these observations are qualitative in actual clinical practice and have not been previously quantified.

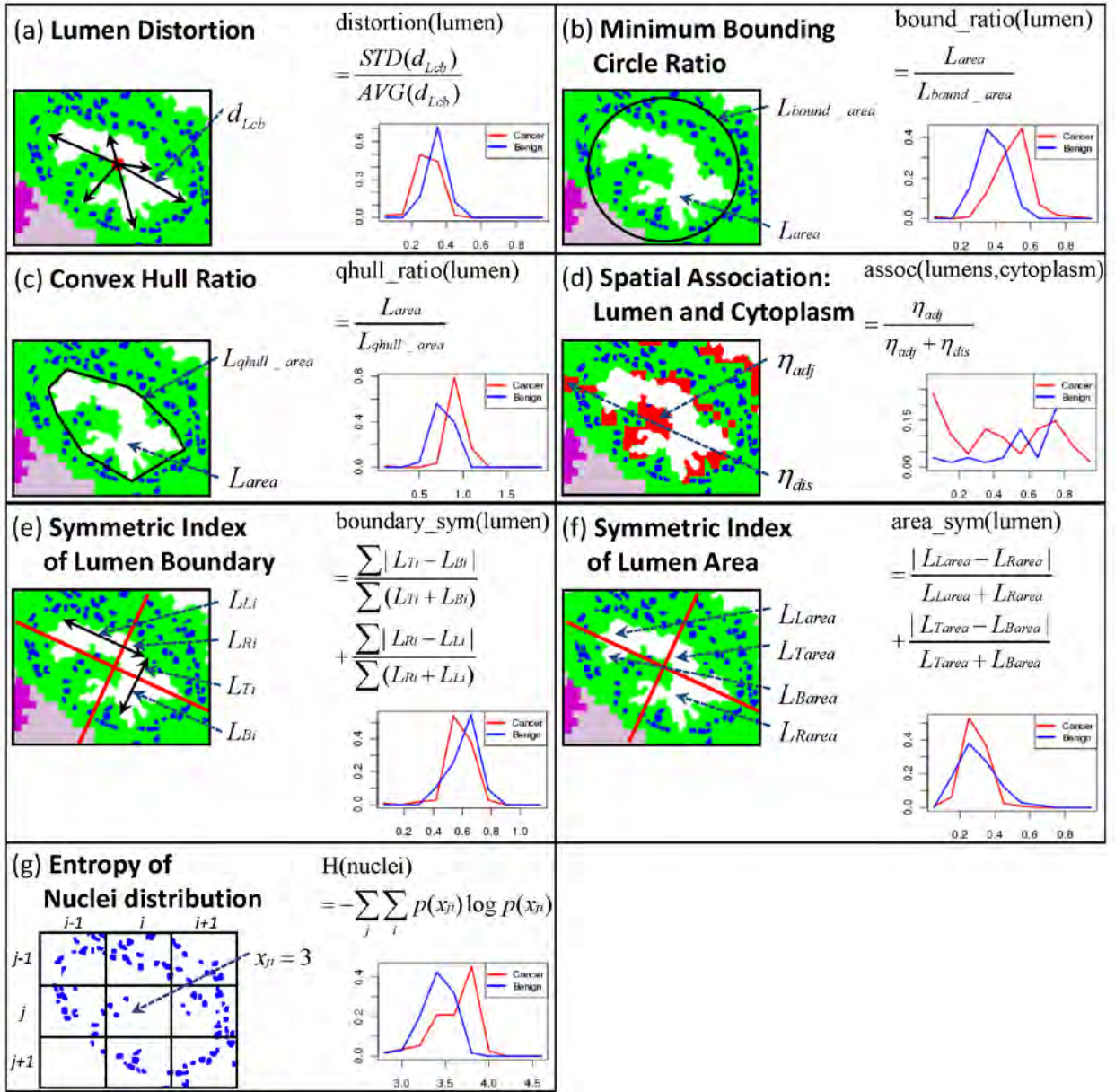


Figure 6. Examples Features.

Each panel shows one example feature, along with the distributions of the feature's values for cancer (red) and benign (blue) classes.

3.1. Epithelial cell-related features

We use epithelial cell type classification from IR data to measure epithelium-related features. However, individual epithelial cells in the tissue are not easily delineated. Therefore, in addition to features directly describing epithelial cells, we also quantify properties of epithelial nuclei, which are available from the segmentation described in §2. The quantities we measure in defining features are: (1) size of epithelial cells, (2) size of epithelial nuclei, (3) number of nuclei in the sample, (4) distance from a nucleus to the closest lumen, (5) distance from a nucleus to the epithelial cell boundary, (6) number of “isolated” nuclei (nuclei that have no neighboring nucleus within a certain distance), (7) number of nuclei located “far” from lumens, and (8) entropy of spatial distribution of nuclei (Figure 6G). Supplementary Materials provide specifics of these measures and their calculation.

3.2. Lumen-related features

Features describing glands have been shown to be effective in PCa classification. Here, we try to characterize lumens and mostly focus on the differences in the shape of the lumens. The quantities we measure in defining these features are: (1) size of a lumen, (2) number of lumens,

(3) lumen “roundness”, defined as $\frac{L_{peri}}{2L_{area}}r$ where L_{peri} is the perimeter of the lumen, L_{area} is the size of the lumen, and r is the radius of a circle of size L_{area} , (4) lumen “distortion” (Figure 6A),

computed as $\frac{STD(d_{L_{cb}})}{AVG(d_{L_{cb}})}$ where $d_{L_{cb}}$ is the distance from the center of a lumen to the boundary of

the lumen and $AVG(\cdot)$ and $STD(\cdot)$ represent the average and standard deviation, (5) lumen “minimum bounding circle ratio” (Figure 6B), defined as the ratio of the size of a minimum bounding circle of a lumen to the size of the lumen, (6) lumen “convex hull ratio” (Figure 6C), which is the ratio of the size of a convex hull of a lumen to the size of the lumen, (7) symmetric index of lumen boundary (Figure 6E, see Supplementary Materials), (8) symmetric index of lumen area (Figure 6F, see Supplementary Materials), and (9) spatial association of lumens and cytoplasm-rich regions (Figure 6D, see Supplementary Materials). Features (3) – (8) are various ways to summarize lumen shapes, while feature (9) is motivated by the loss of functional polarization of epithelial cells in cancerous tissues.

3.3. Global & local tissue features

We have described above the individual measures of epithelium and lumen related quantities that form the basis of the features used by our classification system. Normally, these features have to be summary measures over the entire tissue sample or desired classification area. Hence, we employ average (AVG) or standard deviation (STD), and in some cases the sum total (TOT) of these quantities for further analysis. These features are called “global” features since they are calculated from the entire sample. However, in some cases global features may be misleading, especially where only a part of the tissue sample is indicative of cancer. Therefore, in addition to global features, we define “local” features by sliding a rectangular window of a fixed size (typically 100x100 pixels) throughout a tissue sample, computing the average or sum total of the feature in each window, and computing the standard deviation and/or extrema over the values for

all windows (Figure 7). In all, 67 features (29 global and 38 local features) are defined capturing various aspects of tissue morphology.

4. Feature Selection

Feature selection is the step where the classifier examines all available features (67 in our case) with respect to the training samples, and selects a subset to use on test data. This selection is generally based on the criterion of high accuracy on training data, but also strives to ensure generalizability beyond the training data. We adopt a two-stage feature selection approach here. In the first stage, we generate a set of candidate features ($C_{candidate}$) by using the so-called minimum-redundancy-maximal-relevance (mRMR) criterion. In each iteration, given a feature set chosen thus far, mRMR chooses the single additional feature that is least redundant with the chosen features, while being highly correlated with the class label. $C_{candidate}$ is a set of features that is expected to be close to the optimal feature set for a dataset and a classifier under consideration. It is constructed as follows. Given a feature set $F = (f_1, \dots, f_M)$ ordered by mRMR, AUC of the set of i top-ranked features is computed for varying values of i . We limit the value of i to be ≤ 30 . The feature subset with the best AUC is chosen as the $C_{candidate}$. In the second stage, feature selection continues with $C_{candidate}$ as the starting point, using the sequential floating forward selection (SFFS) method. This method sequentially adds new features followed by conditional deletion(s) of already selected features. Starting with the $C_{candidate}$, SFFS searches for a feature $x \notin C_{candidate}$ that maximizes the AUC among all feature sets $C_{candidate} \cup \{x\}$, and adds it to $C_{candidate}$. Then, it finds a feature $x \in C_{candidate}$ that maximizes the AUC among all feature sets $C_{candidate} - \{x\}$. If the removal of x improves the highest AUC obtained by $C_{candidate}$, x is deleted from $C_{candidate}$. As long as this removal improves upon the highest AUC obtained so far, the removal step is repeated. SFFS repeats the addition and removal steps until AUC reaches 1.0 or the number of additions and deletions exceeds 20, and the feature set with the highest AUC thus far is chosen as the optimal feature set. The classification capability of a feature set, required for feature selection, is measured by the area under the ROC curve (AUC), obtained by cross-validation on the training set.

5. Classification

We note that there are two levels of classification here. In the first, IR spectral data is used to provide histologic images where each pixel has been classified as a cell type. In the second, the measures from H&E images and IR images are used to classify tissue into disease states. In this manuscript, we do not discuss the first classification task as its development and results are well-documented. For the latter task, we used a well established classification algorithm, namely support vector machine (SVM). Two cost factors are introduced to deal with an imbalance in training data. The ratio between two cost functions was chosen as

$$\frac{C_+}{C_-} = \frac{\text{number of negative training examples}}{\text{number of positive training examples}}$$

to make the potential total cost of the false positives and the false negatives the same. (See Supplementary Materials for details.)

6. Data preparation

All of the H&E stained images were acquired on a standard optical microscope at 40x magnification. The size of each pixel is 0.9636 μm x 0.9636 μm . On the other hand, the pixel size of IR images is 6.25 μm x 6.25 μm . The acquisition was previously described in previous

years' reports. Two data sets, stained under different conditions, were used in this study. The first dataset ("Data1") consists of 66 benign samples and 115 cancer samples, and the second set ("Data2") includes 14 benign and 36 cancer samples. These were previously acquired under the grant.

Results and discussion: We then applied the methods to classify prostate tissue and the results are presented below.

1. The classification system achieves AUC greater than 0.97 on both data sets

We first performed K -fold cross validation on each dataset. The data set was divided into K roughly equal-sized partitions, one partition was left out as the "test data", the classifier was trained on the union of the remaining $K - 1$ partitions (the "training data") and evaluated on the test data. This was repeated K times, with different choices of the left-out partition. (We set $K = 10$.) In each repetition, cross-validation on the training data was used to select the feature set with the highest AUC as explained in §4. The correct and incorrect predictions in the test data, across all K repetitions, were summarized into a ROC plot and the AUC was computed, along with specificities when sensitivity equals 90, 95, or 99%. Since the cross-validation exercise makes random choices in partitioning the data set, we examined averages of these performance metrics over 10 repeats of the entire cross validation pipeline. The average AUC for *Data1* and *Data2* were 0.982 and 0.974 respectively (Table 1, "feature extraction" = "IR & HE"). At 90%, 95%, and 99% sensitivities, the average specificity achieved on *Data1* was 94.76%, 90.91%, and 77.80% respectively, while that on *Data2* was 92.53%, 84.19%, and 49.54% respectively.

One way to interpret the above values is to examine our automated pipeline as a pre-screening mechanism to identify the samples to be examined by a human pathologist. At a "true positive rate" of 99% (which means that only 1% of the cancer samples will be missed by the screen), the "false positive rate" is 22.2% (i.e., 22.2% of the benign samples will make it through the screen) on average for *Data1* (Table1), thereby reducing the workload of the pathologist by 4.5-fold. While the error rate of manual pathology determinations is generally accepted to be in 1-5% range, inclusion of confounding cancer mimickers raises the rate to as high as 7.5%. Also noteworthy is the observation that the same algorithm performs consistently well on both data sets, that were obtained from different staining conditions. This speaks to the robustness of the classification framework, an attribute that we investigated further in the next exercise.

2. Classification system is robust to staining conditions

Here, we trained a classifier on *Data1* and tested its performance on *Data2*. We observed an average AUC of 0.956, with average specificity of 88.57%, 81.92%, and 26.86% at sensitivity equaling 90%, 95%, and 99% respectively (Table 2, "feature extraction" = "IR & HE"). These values are competitive with the cross-validation results on *Data2* (Table 1), where the training and testing were both performed on (disjoint parts of) *Data2*.

3. IR data is critical to classification performance

To assess the utility of the IR-based cell-type classification, we repeated the above exercises after extracting features without the guidance of the IR data; i.e., epithelial cells were predicted from the H&E images alone (see Supplementary Materials for details). All of the features

defined in §3 were used, except for “Spatial association of lumens and apical regions”, since the distinction between cytoplasm-rich and nuclear-rich region in epithelial cells was unclear in H&E images. The results from this disadvantaged classifier are shown in Tables 1 and 2 (“feature extraction” = “HE only”). For both types of experiments, we obtained lower average AUCs and specificity values. For instance, the AUC of cross-validation in *Data2* (Table 1) dropped from 0.974 to 0.880. Similarly, the results of validation between datasets (Table 2) were substantially worse now compared to the IR-guided classification, with the AUC dropping from 0.956 to 0.918. This indicates that feature extraction with the help of the IR cell-type classification is critical to consistent and reliable classification of cancer versus benign tissue samples.

Dataset	Feature Extraction	AUC		Sensitivity (%)	Specificity (%)		M _f
		AVG	STD		AVG	STD	
Data1	IR & HE	0.982	0.0030	90	94.76	1.64	13
				95	90.91	1.62	
				99	77.80	5.52	
	HE only	0.968	0.0052	90	91.64	2.26	11
				95	83.90	1.91	
				99	53.43	13.65	
Data2	IR & HE	0.974	0.0145	90	92.53	7.11	7
				95	84.19	10.84	
				99	49.54	22.51	
	HE only	0.880	0.0175	90	61.34	10.31	8
				95	22.21	10.06	
				99	11.21	6.01	

Table 1 . Classification results via cross-validation.

AVG and STD denote average and standard deviation across ten repeats of cross-validation. M_f is the median size of the feature set obtained by feature selection from training data. Column “Feature Extraction” indicates if features were obtained using H&E as well as IR data, or with H&E data alone.

Feature Extraction	Dataset	AUC		Sensitivity (%)	Specificity (%)		M _f
		AVG	STD		AVG	STD	
IR & HE	Train	0.994	0.0006	90	98.30	0.68	13
				95	96.58	1.10	
				99	91.55	2.55	
	Test	0.956	0.0089	90	88.57	5.96	
				95	81.92	5.28	
				99	26.86	15.50	
HE only	Train	0.986	0.0021	90	97.77	0.97	10
				95	91.56	2.49	
				99	79.29	4.47	
	Test	0.918	0.0100	90	65.51	8.37	
				95	46.14	7.53	
				99	13.29	6.94	

Table 2. Validation between datasets.

A classifier is trained on *Data1* and tested on *Data2*. AVG and STD denote the average and standard deviation. M_f is the median size of the optimal feature set. Column “Feature Extraction” indicates if features were obtained using H&E as well as IR data, or with H&E data alone. Column “Dataset” indicates if the performance metrics are from training data (*Data1*) or from test data (*Data2*).

Previously, Tabeshi *et al.* achieved an accuracy of 96.7% via cross validation in cancer/no-cancer classification. Color, morphometric, and texture features were extracted, and all images were acquired under similar conditions. We note that our classification result (Table 1), based solely on morphology, is comparable to their result; however the software developed by Tabeshi *et al.* was not available for evaluation in our data sets. Color and texture features could provide additional information; however, their robustness to different data sets is questionable, and their interpretation is not as obvious as that of morphological features, which are used in clinical practice. Different data sets may have varied properties which may be attributable to staining variations, inconsistent image acquisition settings, and image preparation. The performance of the same method based on texture features has been seen to greatly change from one data set to another. Variations in staining may affect color features. In contrast, morphological features were shown to be robust to varying image acquisition settings. Nonetheless, the quality of morphological features is subject to segmentation of histologic objects. Thus, any method based on morphological features will benefit from the IR cell-type classification.

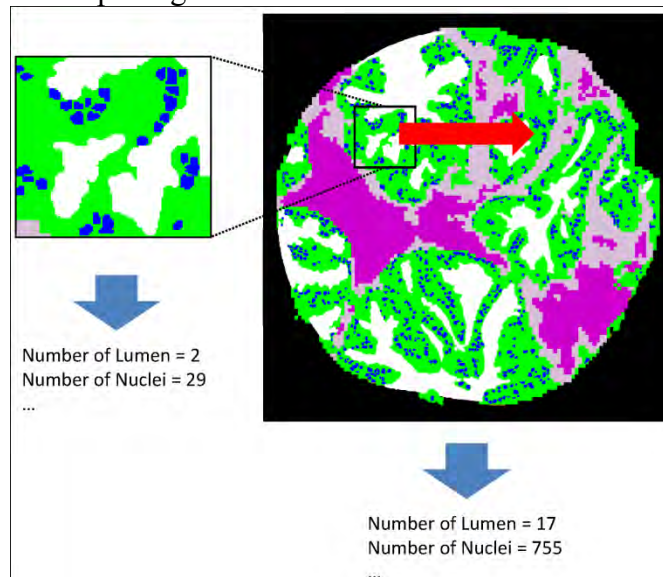


Figure 7. Global and Local Feature Extraction.

Global features are extracted from the entire tissue sample, and local features are extracted by sliding a window of a fixed size across the tissue sample and computing summary statistics, such as standard deviation, of window-specific scores. In this example, the global feature “number of nuclei” has value 755, while one example position of the sliding window is shown, with “number of nuclei” = 29.

4. Examination of discriminative features

We examined the importance of each feature by its rank in the first phase of feature selection, based on its “relevance” to the class label (see Supplementary Materials, mRMR). Since

different features (e.g., average or standard deviation, global or local features) based on the same underlying quantity (e.g., “lumen roundness”) generally have similar relevance, we examined the average relevance of features in each of 17 feature categories (Figure 8), for each data set. The complete list of the individual features and their relevance and mRMR rank (for *Data1*) is available in Figure 9. For *Data1*, lumen-related feature categories are most relevant in general, while epithelium-related feature categories are most important for *Data2*. It is surprising that the top 3 feature categories in *Data1* (Figure 8, blue bars) – size of lumen, lumen roundness, and lumen convex hull ratio – have very low relevance in *Data2*, although we note that this may be in large part due to variations in staining and malignancy of tumors between the two data sets. Also, examining the features (or feature categories) with highest relevance alone may be slightly misleading, because this examination does not account for redundancy among features.

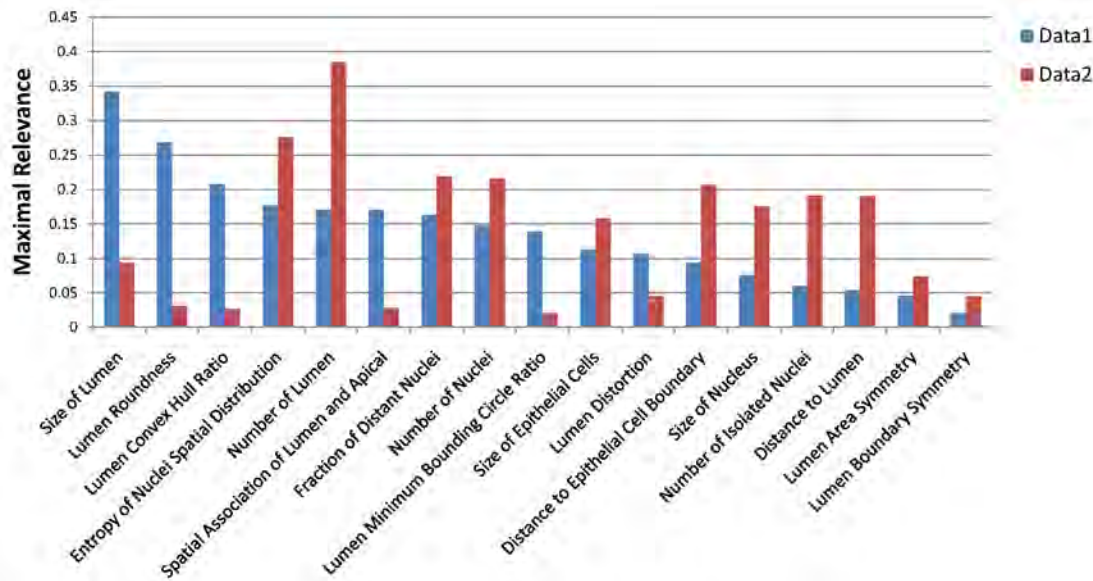


Figure 8. Importance of 17 feature categories.

The average “maximal relevance” of features belonging to each feature category is shown, for both data sets, sorted in decreasing order for the first data set.

Feature Name	Type	Maximal Relevance	mRMR rank
Size of Lumen	G_AVG	0.501	1
Lumen Roundness	G_AVG	0.408	7
Size of Lumen*	L_AVG	0.444	8
Size of Lumen	G_TOT	0.369	12
Lumen Convex Hull Ratio	G_TOT	0.401	3
Lumen Roundness	L_MAX	0.357	9
Lumen Convex Hull Ratio	L_MAX	0.366	16
Size of Lumen	L_MAX	0.354	21
Size of Lumen*	L_TOT	0.35	25
Size of Lumen*	L_MAX	0.339	18
Size of Lumen*	L_MAX	0.314	31
Size of Lumen	L_MAX	0.312	36
Size of Lumen	L_TOT	0.284	46
Size of Lumen	L_MAX	0.255	49
Lumen Roundness	G_TOT	0.234	30
Lumen Minimum Bounding Circle Ratio	G_AVG	0.232	14
Size of Lumen	G_TOT	0.226	42
Number of Lumen	G_TOT	0.225	10
Entropy of Nuclei Spatial Distribution	L_MAX	0.218	6
Entropy of Nuclei Spatial Distribution	G_TOT	0.208	2
Lumen Roundness	L_TOT	0.2	26
Lumen Minimum Bounding Circle Ratio	L_MAX	0.197	39
Size of Nucleus	G_TOT	0.189	23
Number of Nuclei	G_TOT	0.187	40
Distance to Epithelial Cell Boundary	G_TOT	0.18	33
Spatial Association of Lumen and Cytoplasm	G_TOT	0.17	11
Number of Lumen	L_TOT	0.165	4
Size of Nucleus	L_TOT	0.163	19
Fraction of Distance Nuclei	G_TOT	0.163	22
Size of Epithelial Cells	G_TOT	0.159	32
Lumen Distortion	G_AVG	0.146	34
Size of Epithelial Cells	L_MAX	0.143	13
Distance to Lumen	L_MAX	0.143	38
Lumen Distortion	L_MAX	0.131	52
Number of Lumen	L_MAX	0.121	29
Entropy of Nuclei Spatial Distribution	L_TOT	0.105	54
Size of Nucleus	L_MAX	0.103	24
Distance to Epithelial Cell Boundary	L_MAX	0.098	51
Lumen Minimum Bounding Circle Ratio	L_TOT	0.088	17
Number of Isolated Nuclei	G_TOT	0.087	8
Lumen Minimum Bounding Circle Ratio	G_TOT	0.077	37
Symmetric Index of Lumen Area	L_MAX	0.073	41
Symmetric Index of Lumen Area	G_TOT	0.063	20
Lumen Distortion	G_TOT	0.059	27
Distance to Epithelial Cell Boundary	L_MAX	0.059	35
Number of Nuclei	L_MAX	0.057	63
Distance to Lumen	G_TOT	0.053	62
Number of Isolated Nuclei	L_MAX	0.051	28
Symmetric Index of Lumen Boundary	L_TOT	0.051	47
Lumen Convex Hull Ratio	G_TOT	0.046	65
Symmetric Index of Lumen Area	G_TOT	0.043	50
Lumen Distortion	L_TOT	0.043	53
Symmetric Index of Lumen Boundary	G_TOT	0.042	33
Distance to Epithelial Cell Boundary	G_TOT	0.039	45
Size of Epithelial Cells	L_TOT	0.038	43
Size of Nucleus	L_MAX	0.037	48
Lumen Convex Hull Ratio	L_TOT	0.03	56
Size of Nucleus	G_TOT	0.023	44
Symmetric Index of Lumen Area	L_TOT	0.019	55
Symmetric Index of Lumen Boundary	L_MAX	0.019	58
Symmetric Index of Lumen Boundary	G_TOT	0.018	61
Distance to Lumen	L_MAX	0.018	64
Size of Nucleus	G_TOT	0.014	59
Size of Nucleus	L_TOT	0.008	60
Number of Nuclei	L_TOT	0.006	57
Number of Isolated Nuclei	L_TOT	0.006	66
Distance to Lumen	G_TOT	0.002	67

Figure 9. List of features and their maximal relevance and “mRMR rank”.

In the second column, *G* and *L* represent global and local features, respectively. *AVG*, *STD*, *TOT*, and *MAX* denote the average, standard deviation, total amount, and extremal value of features. * In computing local features representing “size of lumen”, two options are available: one is to consider only the part of the lumen within the window, and the other is to consider the entire lumen into account. Asterisk indicates that the former option was chosen.

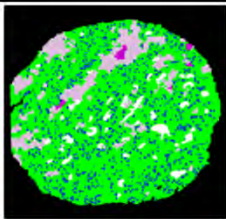
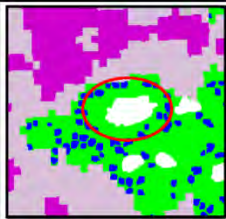
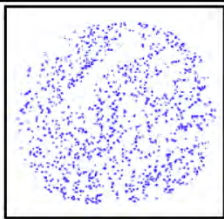
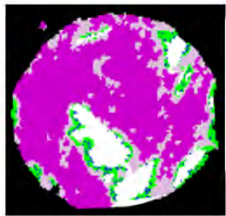
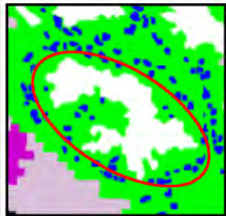
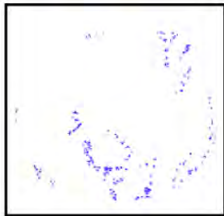
		Number of Lumens Total Size of Nuclei (μm^2)			Roundness		Entropy of Nuclei Distribution
Cancer		45	30094		1.07		3.79
Benign		9	4320		2.21		3.04

Figure 10. Optimal features for distinguishing cancer and benign tissue samples.
The four features shown here are always present in the optimal feature set chosen by the classifier.

Conclusions

In completing this task, we have presented a means to eliminate epithelium recognition deficiencies in classifying H&E images for presence or absence of cancer. The method is entirely transparent to a user and does not involve any adjustment or decision-making based on spectral data. We were able to achieve very effective fusion of the information from two different modalities, namely optical and IR microscopy, that provide very different types of data with different characteristics. Several features of the tissue were quantified and employed for classification. We found that robust classification could be achieved using a few measures, which are detailed to arise from epithelial/lumen organization and provide a reasonable explanation for the accuracy of the model. The choice of combining the IR and optical data is shown to be necessary for achieving the high accuracy values observed. We anticipate that the combined use of the two microscopies – structural and chemical – will lead to an accurate, robust and automated method for determining cancer within biopsy specimens.

TASK 2F: DEVELOP CALIBRATION FOR PREDICTING CANCER GRADE (MONTHS 18-22)

Motivation:

Quality assurance in clinical pathology plays a critical role in the management of patients with prostate cancer as pathology is the gold standard of diagnosis and forms a cornerstone of patient therapy. Methods to integrate quality development, quality maintenance, and quality improvement to ensure accurate and consistent test results are, hence, critical to cancer

management in any setting. These factors have a direct bearing on patient outcomes, financial aspects of disease management as well as malpractice concerns. One of the major failings in prostate pathology today is the rate of missed tumors and variability in grading. It is well known that the grading of prostate tissues suffers from intra- and inter-pathologist variability. In the studies of intra- and inter-pathologist reproducibility, the exact intra-pathologist agreement was achieved in 43-78% of the instances, and in 36-81% of the instances, the exact inter-pathologist agreement was reported. It is also known that the variability of the grading could be reduced after pathologists are re-trained. There could be many ways to educate pathologists such as meetings, courses, online tutorials, and etc, but these are not time- and cost-effective for routine everyday decisions. Therefore, building an automated, fast, and objective method to aid pathologists to examine prostate tissues will greatly help to attain reliable and consistent diagnoses. This will reduce healthcare costs and the chances of malpractice lawsuits as well as improve patient outcomes in therapy.

Innovation in our approach and potential benefits:

When a pathologist examines tissue, he/she looks at a stained imaged of tissue and mentally compares it against a database of previous knowledge or information in books. In essence, the pathologist is manually matching structural patterns he/she has seen earlier and mentally recalling the diagnosis made such that he/she can make the same diagnosis in the specific test case. Here, we report developing a computer information and management and decision-making system that relies of one or more measures of the structure of tissue to provide images from a database that are similar to the sample under consideration. We emphasize that the system does not provide a diagnosis but simply provides the closest matching cases that enable a pathologist to make a diagnosis. We also propose here the new idea of constructing a database of pre-examined prostate tissues and providing similar tissue samples with pathologists from the database while they examine an unknown tissue sample. To our knowledge, no such system currently exists. Further, we propose that our system may or may not use infrared chemical imaging data in comparisons. Comparing with the pre-examined tissues samples, we expect that pathologists to make more consistent and accurate decision. As we build a database of prostate tissue samples, we represent each tissue sample by its morphology. Given an unknown tissue sample, the similarities between the unknown sample and the tissue samples in the database are measured based on the morphological properties, and the most similar tissue samples are retrieved. The pathologist may indicate that certain matches were better than others, resulting in an updating of the database and matching algorithms as needed. The updating may be conducted in real-time.

Work accomplished:

Morphological features have been shown to be able to characterize prostate tissues and can be used for the diagnostic purpose. Here, 67 morphological features, which are based on lumens and epithelial nuclei, were extracted from each tissue sample. The database stores the morphological features for the tissue samples which have already been examined by pathologists.

Once we have an unknown prostate tissue sample (query), first of all, the morphological features are extracted from the tissue sample. Secondly, the similarities between the query and the tissue samples in the database are computed using Euclidean distance based on the morphological features. Lastly, the most similar k tissue samples to the query are retrieved from the database.

To assess the goodness of the method, we have tested our method on a dataset composed of 181 tissue samples. In the dataset, 5, 23, 66, and 21 tissue samples are Gleason grade 2, 3, 4, and 5 cancer (“*Cancer*”), respectively, and 20 and 46 tissue samples are BPH and normal (“*Benign*”), respectively. Due to the small number of tissue samples, Gleason grade 2 is ignored for the further consideration. As mentioned above, each of tissue samples is represented by 67 morphological features.

In order to measure the performance of the method, we adopted k -nearest neighbor (kNN) algorithm and predicted the grade of the query by majority voting. Both accuracy and kappa-coefficient were computed for the predictions. Since pathologists may be more interested in grading of cancerous tissue samples, we also applied our method only to the “*Cancer*” tissue samples; i.e., Gleason grade 3, 4, and 5 samples.

We performed Leave-one-out (LOO) cross-validation on the dataset. LOO leaves one example as a validation data and uses the remaining examples as training data. In our method, the validation data is the query, and the training data is regarded as the database. It should be noted that the number of tissue samples in each grade in the dataset varies. The imbalance in the dataset could affect the prediction made by kNN algorithm. To tackle the problem, we randomly selected the same number of tissue samples from each grade and performed LOO on the sub-dataset. This repeated 100 times, and the average accuracy and kappa-coefficient were computed over the repeats.

Our method is subject to the choice of the number of nearest neighbors to consider for the prediction and the number of features to use for the similarity computation. To examine the effect of them, we computed the average accuracy and kappa-coefficient over 100 repeats as increasing the two factors (Fig. 1). The accuracy decreases as increasing the number of nearest neighbors, and the more features we use, the higher accuracy achieved. The highest average accuracy achieved for grading both “*Cancer*” and “*Benign*” samples (i.e., 5 grades) was 42% using 7 features and 1 nearest neighbor (Fig. 1a). By using 8 features and 1 nearest neighbor, the highest accuracy of 52% achieved for grading only “*Cancer*” samples (i.e., 3 grades) (Fig. 1c). Both cases also achieved the average Kappa coefficient of 0.27 (Fig. 1b, d). In Fig. 2, the distribution of the grade of the retrieved samples is shown. Distinction between “*Cancer*” and “*Benign*” samples is obvious (Fig 2a), but among “*Cancer*”, the retrieved samples often do not belong to the same grade with the query, especially between Gleason grade 3 and 4.

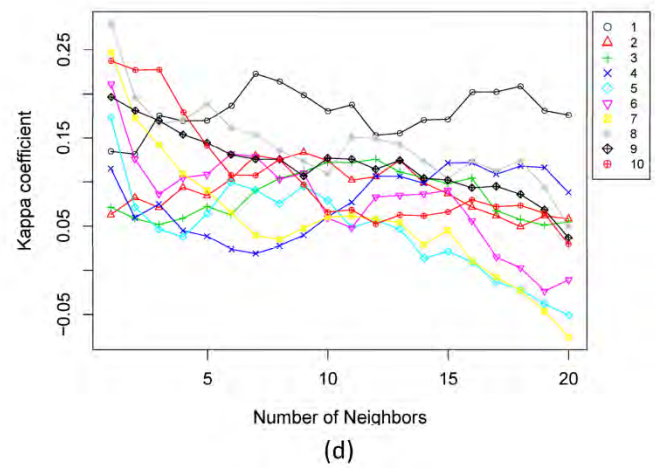
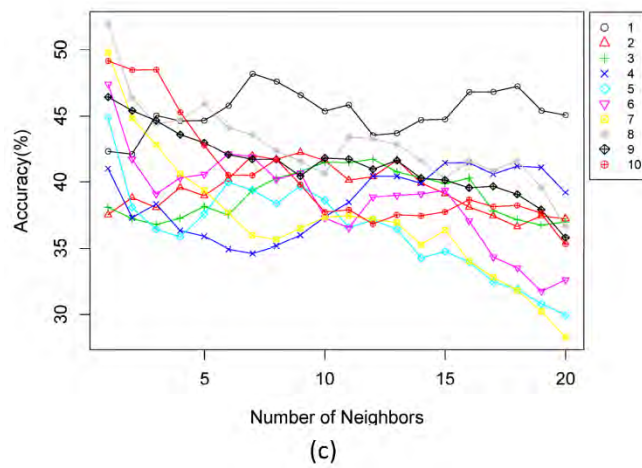
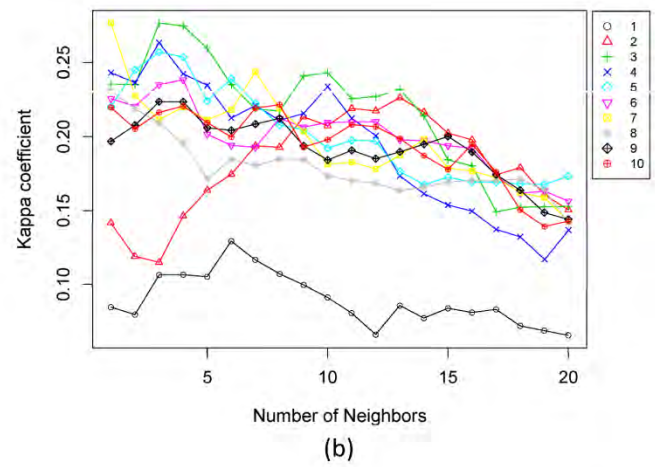
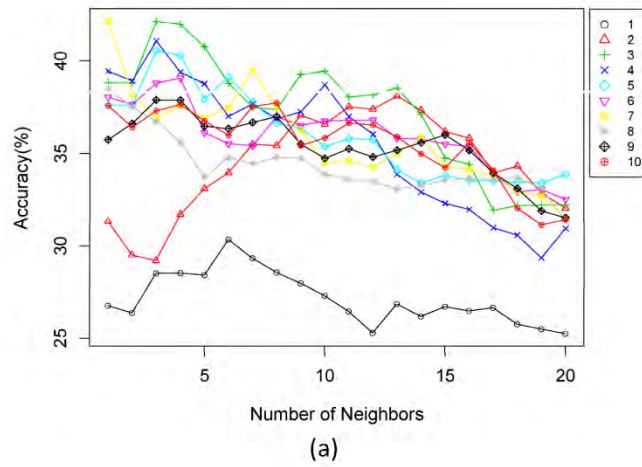


Figure 11. Average accuracy and kappa coefficient. (a), (b) grading for both “*Cancer*” and “*Benign*” samples. (c), (d) grading for “*Cancer*” samples. Each line depicts the accuracy and kappa coefficient values of the corresponding number of features.

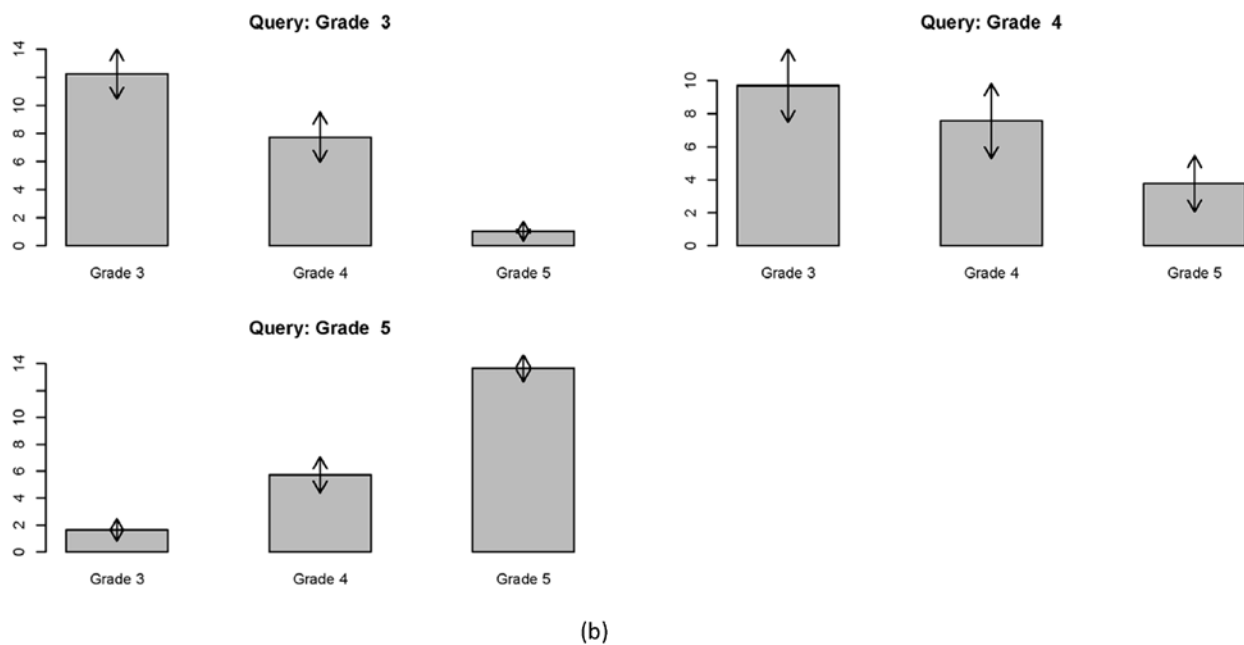
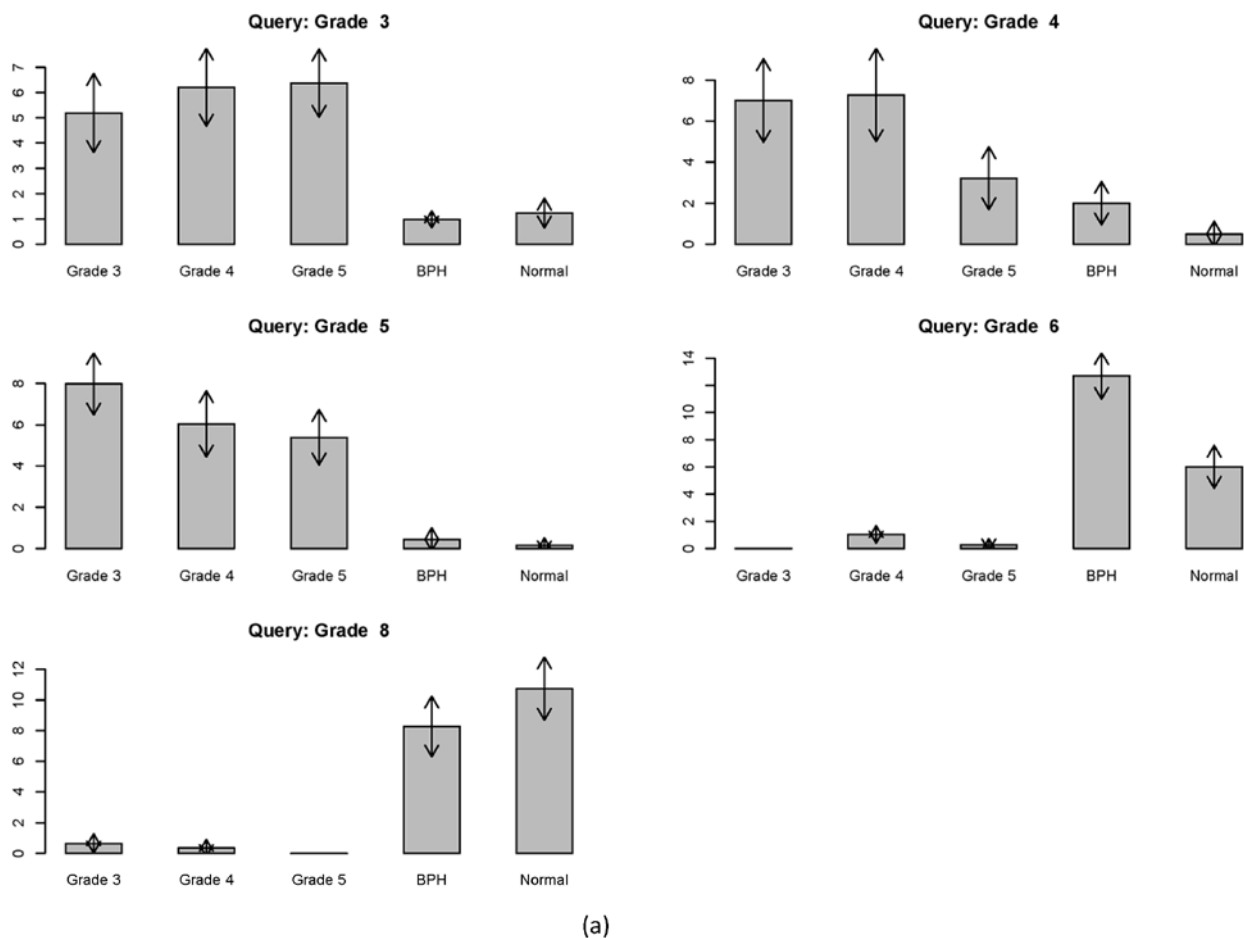


Figure 12. Distribution of the grade of the retrieved samples. (a) grading for both "Cancer" and "Benign" samples. (b) grading for "Cancer" samples. For the samples in each grade, the

grade of retrieved samples are counted and the average number of samples are shown. The arrows denote ± 1 standard deviation of the number of samples.

TASK 2G: DEVELOP PROTOCOLS AND VALIDATE GLEASON GRADING OF TUMOR (MONTHS 18-27)

The task above provides details of the development and LOO validation. More rigorous validations are needed but the preliminary results shown here have been used to validate the grading correspondence and the protocols we have developed, as noted above.

It is important to place the magnitude of our advance in context. Several research efforts have been made to develop automated systems for the grading of prostate tissues. The majority of systems have been used texture and/or morphological features to characterize and classify tissue samples into correct classes. However, the information which pathologists will obtain by using such methods may be limited since these only provide the predicted grade in general. The prediction also relies on the training data. Most importantly, these prior efforts always sought to match a sample completely to provide a diagnosis, rather than provide matching candidates. Further, the role of other modalities in the process was not clear. Here, we may also use IR chemical imaging data in matching. Our premise is that tissue samples which have the same grade and similar characteristics and patterns with the sample of interest will afford more information to pathologists and hence, the system enables a matching to a database rather than seeking to provide an unequivocal diagnosis.

Future outlook enabled by this progress:

The matching system would be implemented first for a clinical trial and then, would be ready for commercial translation. While a true clinical trial is the next step, some further development of the actual methods may be expected. We have built the method into existing software as a user-friendly software.

Task 3. Develop mathematical framework to correlate spectral, spatial and clinical parameters with cancer progression

- a. Identify and validate spectral metrics and develop spatial metrics indicative of tumor grade (Months 27-30)
- b. Develop prediction algorithm for predicting outcome (Months 30-36)

Activities:

We have imaged 460 patients with full outcome data and identified several metrics that are indicative of tumor grade (please see task 2 as well). A mathematical framework for correlate the spectral, spatial and clinical parameters with cancer progression has been built using logistic regression. The prediction algorithm is available for use and will be validated. The task for this project was to develop the algorithms and, hence, the task is complete.

"Vcunl60Y tksgr qtwl'cpf 'hpcrk g'cni qtkj o u'lpv'ulhw ctg"O qpj u'55/58+"
"

"A number of reports (invention disclosure, conference etc.) have been written and manuscripts based on this work have been submitted and have been printed as detailed in the following sections.

In summary, the promised work has been accomplished to a reasonable degree and has opened up significant doors to future progress in prostate pathology as a research direction as well as for patients and clinicians.

"
Mg{ 'Tgugtej 'Ceeqo r rkuj o gpw'E'
"

C'i gpgvle'cni qtkj o 'dcugf "'o gj qf 'v'f'kukpi vkuj 'dgpki p'itqo 'o crki pcpv'gr kj grkwo 'wukpi 'kplctgf "
ur gevqueqr le'ko ci kpi 'f'cvc'y cu'uj qy p'v'q'dg'ghgevxg0Ncti g'uecrg'xcrkf'cvkqp'uj qy u'r tqo kulpi 'tguwmu'cpf "
c'o cpwuetkr v'ku'dgkpi 'y tkwgp0'
"

E'Y g'f'vgto kpgf 'y cv'qpg'qh'y g'ng{ 'hcevtu'lp'wpf gtucpf kpi 'qwt 'f'cvc'y cu'y g'ur cvkcn'utwewtg'qh'y g'vwuwg."
""'y cv'enqugn' 'chhgev'f'y g'KT 'f'cvc0C'ugtkgu'qh'uko wrcvkpu'y gtg'eqpf wevgf 'chgt'f'gxgnr kpi 'c'tki qtqwu'qr vlecn
""'o qf gn'v'rtgf lev'f'kvqtvkpu0Tguwmu'ctg'tgr qtv'f'lp'y q'o cpwuetkr v'lp'Cpcr0Ej go 0
"

'EC'eqo dlpcvkqp'qh'KT'cpf 'eqpxgpvkpcnr'cvj qmji { 'ko ci kpi 'j cu'dggp'f'gxgnr gf 'cpf 'gzvgpukxgn' 'xcrkf'cvgf 0'
""'Vj g'o cpwuetkr v'j cu'dggp'uwdo kwgf 'v'q'DO E'ecpegt0
"

'EC'o gj qf 'v'q'eqttgrvg'I rguqp'i tcf gu'y kj 'o gcuwtgf 'f'cvc'j cu'dggp'f'gxgnr gf 0Ncti gt'xcrkf'cvkqp'uwf kgu
""'ctg'pggf gf 0Cp'lp'xgpvkqp'f'kuenquwtg'j cu'dggp'hkrgf 'cpf 'c'r'cvgpv'y kn'dg'hkrgf 'uqpp0Uwdugs wgpv'f'.'y g'y km'
""'uwdo k'c'o cpwuetkr v'dcugf 'qp'y g'y qtn'v'Ecpegt'Tgugtej 0"

Reportable Outcomes.....

Manuscripts

Peer reviewed manuscripts published

1. M.J. Walsh, M.V. Schulmerich, **R. Bhargava** “Progress, critical challenges and a roadmap for translating infrared spectroscopic imaging for cancer histopathology” *Chem. Rev.*, To be submitted (2010 – Invited)
2. J.T. Kwak, S.M. Hewitt, S. Sinha, **R. Bhargava** “Multimodal microscopy for automated histologic analysis of prostate cancer” *BMC Cancer*, Under review (2010)
3. R.K. Reddy, **R. Bhargava** “Automated noise reduction for accurate classification of tissue from low signal-to-noise ratio imaging data” *Analyst*, Under Review (2010)
4. R. Kong, R.K. Reddy, **R. Bhargava** “Characterization of Tumor Progression in Engineered Tissue using Infrared Spectroscopic Imaging” *Analyst* In press (2010)
5. M.J. Nasse, M.J. Walsh, R. Reininger, E.C. Mattson, A. Kujdacsy-Balla, V. Macias, **R. Bhargava**, C.J. Hirschmugl “Fourier Transform Infrared Spectroscopic Imaging with a Synchrotron for Label-free Imaging in Prostate Histology” *Nat. Photonics.*, Under revision for editorial resubmission (2010)
6. B.J. Davis, P.S. Carney, **R. Bhargava** “Theory of mid-infrared absorption microspectroscopy. II. Heterogeneous samples” *Anal. Chem.*, **82**, 3487–3499 (2010) DOI: 10.1021/ac902068e
7. B.J. Davis, P.S. Carney, **R. Bhargava** “Theory of mid-infrared absorption microspectroscopy. I. Homogeneous samples” *Anal. Chem.*, **82**, 3474–3486 (2010) DOI: 10.1021/ac902067p

Book Chapters

1. M.J. Walsh, **R. Bhargava** “Infrared spectroscopic imaging: an integrative approach to pathology” G. Popescu, ed. “Nanobiophotonics” McGraw-Hill (2010 – anticipated)
2. R.K. Reddy, **R. Bhargava** “Chemometric methods for biomedical Raman spectroscopy and imaging” M.D. Morris, P. Matousek, eds. “Emerging Raman Applications and Techniques in Biomedical and Pharmaceutical Fields”, Springer-Verlag, Berlin Heidelberg (2010)
3. A.K. Kodali, **R. Bhargava** “Nanostructured Probes to Enhance Optical and Vibrational Spectroscopic Imaging for Biomedical Applications”, Y.Y. Fu and A. Narlikar, eds. “The Oxford handbook of Nanoscience and Technology: Vol. III”, Oxford University Press, Oxford, UK (2010)

Other manuscripts

1. M.J. Walsh, M.J. Nasse, F.N. Pounder, V. Macias, A. Kujdacsy-Balla, C. Hirschmugl, **R. Bhargava** "Synchrotron FTIR Imaging For The Identification Of Cell Types Within Human Tissues" AIP Conference Proc. Vol. 1214 WIRMS 2009 5th international workshop on infrared microscopy and spectroscopy with accelerator based sources pp. 105-107
2. F.N. Pounder, R.K. Reddy, M.J. Walsh, **R. Bhargava** "Validating the cancer diagnosis potential of mid-infrared spectroscopic imaging" Proc. SPIE 7186, art. no. 71860f
3. **R. Bhargava**, B.J. Davis, "Histologic models for optical tomography and spectroscopy of tissues" Proc. SPIE 7174, 71742H (2009), DOI:10.1117/12.810119

Presentations

Invited conference presentations

First author is the presenting author; First author is also the invited author unless indicated by *

1. **R. Bhargava** "Systems pathology with infrared spectroscopic imaging" Pacificchem 2010, Honolulu, HI, December 2010
 2. **R. Bhargava**, M.V. Schulmerich, A.K. Kodali, R.K. Reddy and M.J. Walsh "Chemical imaging for automated histopathology", Eastern Analytical Symposium, Somerset, November 2010
 3. **R. Bhargava** "Chemical microscopy for systems pathology" Cancer colloquia V II, St. Andrews, Scotland, November 2010
 4. **R. Bhargava**, M.J. Walsh, M.V. Schulmerich "Infrared spectroscopic imaging for label-free and automated histopathology" Frontiers in Optics, Rochester, October 2010
 5. **R. Bhargava**, A. K. Kodali, M. V. Schulmerich, X. Llorca and R. K. Reddy "Integrating physics with chemometrics for enhanced vibrational spectroscopic imaging", FACSS 10, Raleigh, October 2010 [Meggers award symposium]
 6. **R. Bhargava**, R.K. Reddy, A.K. Kodali "Modeling distortions in infrared spectroscopic imaging", FACSS 10, Raleigh, October 2010
 7. **R. Bhargava** "Enabling prostate pathology with infrared spectroscopic imaging – a roadmap for clinical translation", SPEC2010, Manchester, June 2010
 8. **R. Bhargava** "Non-perturbing cancer diagnostics using infrared spectroscopic imaging", Pittcon 2010, Orlando, March 2010
 9. **R. Bhargava** "Progress towards cancer pathology using infrared spectroscopic imaging", Pittcon 2010, Orlando, March 2010
 10. **R. Bhargava** "Enabling systems pathology by infrared spectroscopic imaging", Pittcon 2010, Orlando, March 2010
 11. **R. Bhargava** "Pathology without pathologists?" Pathological Society of Great Britain and Ireland, London, January 2010
 12. M. Schulmerich, A. Kodali, B. Davis, L. J. Elgass, **R. Bhargava***, "Raman Microscopy of layered materials for improved localization of the Raman signal", FACSS 2009, Louisville, October 2009
 13. **Bhargava**, R.K. Reddy, M. Schulmerich, A.K. Kodali, F.N. Pounder B.J. Davis "Next-generation infrared imaging for biomedical spectroscopy", FACSS 09, Louisville, October 2009
 14. **R. Bhargava**, J. Ip, A.K. Kodali, F.N. Pounder B.J. Davis "Ultrafast IR imaging for Biomedical applications", ICAVS-5, Melbourne, July 2009 (*Plenary Lecture*)
-

"

1. Department of Bioengineering, Ohio State University, 2009
2. Beckman Institute Director's Seminar Series, UIUC 2009

Contributed presentations

*First author is the presenting author, unless indicated by **

1. A.K. Kodali, **R. Bhargava** "Designing ultrasensitive SERS-based probes via optimally tailored nanostructures", FACSS 2009, Louisville, October 2009.
2. R.K. Reddy, **R. Bhargava** "Modeling, Data Visualization and Histopathology using Fourier Transform Infrared (FT-IR) Spectroscopic Imaging of Human Tissue Specimens", BMES 2009 Pittsburgh, PA, October 2009
3. M. J. Walsh, M. J. Nasse, F. N. Pounder, V. Macias, A. Kajdacsy-Balla, C. Hirschmugl, **R. Bhargava** "Mid-infrared spectroscopic imaging of prostate tissue towards cancer diagnosis and prognosis", BMES 2009, Pittsburgh, October 2009
4. M. J. Walsh, M. J. Nasse, F. N. Pounder, V. Macias, A. Kajdacsy-Balla, C. Hirschmugl, **R. Bhargava** "Synchrotron FT-IR imaging for identification of cell types within human tissues", WIRMS 2009, Banff, Canada, September 2009
5. L. Feng, **R. Bhargava**, I. Jasiuk "Age-Related Changes in Structure, Composition and Mechanical Properties of Swine Cortical Bone", Summer Bioengineering Conference, Lake Tahoe, CA, June 2009,
6. A.K. Kodali, **R. Bhargava** "Nanospherical aggregates for high throughput and ultrasensitive Raman spectroscopic imaging", PITTCO 2009, Chicago, March 2009.

"Go r m{ o gpv"qt"tgugetej "qrr qtwpkkgu"cr r dgf "hqt"cpf lqt"tgegkxgf "dcugf "qp"gzr gtlkpegtclclpi
"uwr r qtvgf "d{ "j ku"cy ctf 0"
"

Dr. Brynmor Davis, a post-doctoral fellow working on this project obtained employment with Creare Inc., NH.

Conclusion.....

The work accomplished demonstrates clear potential and protocols for classifying prostate tissue. If the protocols are validated in on-going larger studies and translated to the clinic, a new tool for prostate histopathology will be available for pathologists and benefits will be realized by patients.

So What Section

An automated method to assist prostate pathologists is available and can rapidly determine the presence of cancer in biopsies. An automated aid to grading is available to aid pathologists in making accurate decisions. Clinical translation of these discoveries can directly improve prostate healthcare, resulting in better treatment of individuals.

References.....

- ¹ A Jemal, R Siegel, E Ward, T Murray, J Xu, C Smigal, MJ Thun Cancer statistics, 2006 *CA Cancer J Clin* **56**, 106-130 (2006).
- ² SM Gilbert, CB Cavallo, H Kahane, FC Lowe Evidence suggesting PSA cutpoint of 2.5 ng/mL for prompting prostate biopsy: Review of 36,316 biopsies. *Urology* **65**, 549-553 (2005).
- ³ PF Pinsky, GL Andriole, BS Kramer, RB Hayes, PC Prorok, JK Gohagan, Prostate, Lung, Colorectal and Ovarian Project Team Prostate Biopsy Following a Positive Screen in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial *J Urol* **173**, 746-750 (2005). discussion 750-751.
- ⁴ PA Humphrey *Prostate Pathology* American Society of Clinical Pathology, Chicago (2003).
- ⁵ EN Lewis, PJ Treado, RC Reeder, GM Story, AE Dowrey, C Marcott, IW Levin Fourier transform spectroscopic imaging using an infrared focal-plane array detector *Anal. Chem.* **67**, 3377-3384 (1995).
- ⁶ R Bhargava, SQ Wang, JL Koenig Processing FTIR Imaging Data for Morphology Visualization *Appl Spectrosc* **54**, 1690-1706 (2000).
- ⁷ DC Fernandez, R Bhargava, SM Hewitt, IW Levin Infrared spectroscopic imaging for histopathologic recognition *Nat. Biotechnol.* **23**, 469-474 (2005).

APPENDICES - NEXT PAGE.

Towards a practical Fourier transform infrared chemical imaging protocol for cancer histopathology

Rohit Bhargava

Received: 30 May 2007 / Revised: 9 July 2007 / Accepted: 12 July 2007 / Published online: 5 September 2007
© Springer-Verlag 2007

Abstract Fourier transform infrared (FTIR) chemical imaging is a strongly emerging technology that is being increasingly applied to examine tissues in a high-throughput manner. The resulting data quality and quantity have permitted several groups to provide evidence for applicability to cancer pathology. It is critical to understand, however, that an integrated approach with optimal data acquisition, classification, and validation is necessary to realize practical protocols that can be translated to the clinic. Here, we first review the development of technology relevant to clinical translation of FTIR imaging for cancer pathology. The role of each component in this approach is discussed separately by quantitative analysis of the effects of changing parameters on the classification results. We focus on the histology of prostate tissue to illustrate factors in developing a practical protocol for automated histopathology. Next, we demonstrate how these protocols can be used to analyze the effect of experimental parameters on prediction accuracy by analyzing the effects of varying spatial resolution, spectral resolution, and signal to noise ratio. Classification accuracy is shown to depend on the signal to noise ratio of recorded data, while depending only weakly on spectral resolution.

Keywords Fourier transform infrared spectroscopy · FTIR imaging · Infrared microscopy · Prostate · Histopathology · Microspectroscopy

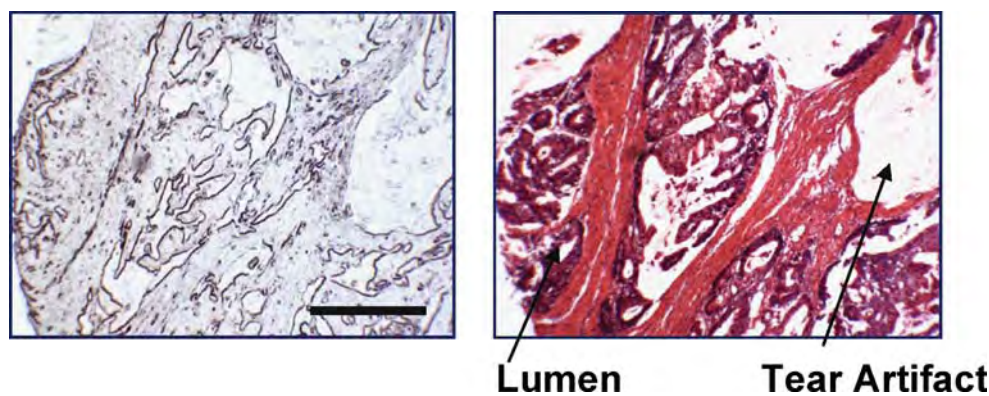
Introduction

Cancer is one of the leading causes of death in the western world and is becoming increasingly prevalent worldwide. It is well established that appropriate therapy for cancers diagnosed early generally leads to improved prognosis and longer survival. Consequently, population screening tests to detect disease are increasingly being deployed. The emphasis in screening populations is on obtaining a high sensitivity through simple diagnostic tests. For example, the prostate-specific antigen (PSA) assay [1] helps triage persons at risk for prostate cancer. A cutoff level (typically 4 ng mL^{-1}) or increase in PSA velocity implies that the screened person should be at heightened surveillance and typically undergoes a biopsy to confirm disease. Morphologic structures in biopsied tissue, as diagnosed by a pathologist, are the only definitive indicator of disease and form the gold standard of diagnosis [2]. Along with clinical history, stage, and PSA values, pathologic diagnoses form a cornerstone of clinical therapy and serve as a basis for a vast majority of research activity [3].

Typically, multiple samples are withdrawn from the organ during biopsy. Extracted tissue samples are fixed, embedded, and sectioned (typically to 1- to 5- μm thickness) onto a glass slide for review. By itself, tissue does not have much useful contrast in optical brightfield microscopy. Hence, the prepared slide is stained with dyes. A mixture of hematoxylin and eosin (H&E) is commonly employed, staining protein-rich regions pink and nucleic acid-rich regions of the tissue blue, for example, as shown in Fig. 1. Using the contrast, a trained person can recognize specific cell types and alterations in local tissue morphology that are indicative of disease. In prostate tissue, epithelial cells line three-dimensional ducts. In two-dimensional thin sections, thus, the cells appear to line empty circular regions (lumen).

R. Bhargava (✉)
Department of Bioengineering and Beckman Institute for
Advanced Science and Technology,
University of Illinois at Urbana-Champaign,
Urbana, IL 61801, USA
e-mail: rxb@uiuc.edu

Fig. 1 Brightfield microscopy images of unstained (*left*) and stained (*right*) prostate tissue sections. Hematoxylin and eosin (H&E) stains provides contrast, allowing a trained person to recognize epithelial cells and ductal structure (lumen), while ignoring artifacts and confounding morphologies. A trained human can also learn to robustly recognize patterns within lumen that indicate cancer. The *scale bar* corresponds to 100 μm



Distortions in normal lumen appearance provide evidence of cancer and characterize its severity (grade). The process is fundamentally a manual pattern recognition that seeks to match observations to known healthy or diseased morphologies.

Manual examination of biopsies is very powerful in that humans can not only recognize disease generally but can also overcome confounding preparation artifacts, detect unusual cases, and recognize deficiencies in diagnostic quality. This capability of considering and neglecting features based on prior knowledge is crucial for accurate and robust diagnoses. The process, however, is time consuming, allows for limited throughput and, frequently, leads to variance in subjective judgments about the disease severity, i.e., grade [4]. As an alternative, computer-based pattern recognition approaches to diagnose disease may provide more accurate, reproducible, and automated approaches that could reduce variance in diagnosis while proving economically favorable. Hence, attempts have been made to characterize morphology using H&E image analysis as well as biomarkers to stain for specific molecular features. Automated approaches that can rival human performance in usual clinical settings, however, are still unavailable. Specifically, the attributes of high accuracy and robust applicability are lacking.

The information content of H&E-stained images is limited and attempts to automatically recognize structural patterns indicative of prostate cancer, unfortunately, have not led to clinical protocols. Similarly, probe-based molecular imaging can provide exquisite information regarding the location and content of specific epitopes but is limited by complex diseases not expressing universally the same epitopes or panels of markers. Stains used can generally detect one feature that may aid diagnosis (e.g., AMACR) but do not provide entire diagnostic information in themselves. An exciting alternative is emerging in the form of chemical imaging and microscopy [5]. As opposed to conventional dye-assisted imaging or probe-assisted molecular imaging, chemical imaging [6] seeks to directly measure the identity and/or concentration of chemical species in the sample using spectroscopy. Hence, no

molecular probes (MPs) are needed to see the presence of specific epitopes; instead computer algorithms are used to extract information from the data (instead of MP hybridization) and statistical methods are used to provide confidence (as opposed to brown tints for MPs). The approach is limited only by the ability of the technology to sense specific types of molecules or otherwise resolve chemical species and morphologic structures. Among the prominent approaches are vibrational spectroscopic imaging, both Raman and infrared (IR), as well as mass spectroscopic imaging (MSI) [7, 8] and magnetic resonance spectroscopic imaging (MRSI) [9]. While each technology promises a specific measurement (e.g., proteins or metabolic products) for specific situations (e.g., in vivo or ex vivo), IR spectroscopic imaging [10] is particularly attractive for the analysis of tissue biopsies in that it permits a rapid and simultaneous fingerprinting of inherent biologic content, extraneous materials, and metabolic state [11–14].

IR spectroscopic imaging, generally practiced using interferometry and termed Fourier transform infrared (FTIR) spectroscopic imaging or, succinctly, FTIR imaging, offers a particular combination of spatial, spectral, and chemical detail [15]. Limitations of FTIR imaging include coarser spatial resolution compared to Raman imaging or high powered optical microscopy and lack of specific molecular detail compared to MSI. Tissue biopsies are examined as thin sections on a solid substrate. The tissue is dehydrated and is stable due to fixation. Typically, structures of pathologic interest are several to hundreds of micrometers in size, requiring fairly moderate magnifications for decision making. These conditions imply that the need to image in vivo, at exceptionally high spatial resolution, or in aqueous environments is not critical and that standard pathologic laboratory processing can be employed for IR imaging. Due to the linear absorption process being utilized, the signal from IR spectroscopy is large and readily obtained, promising relatively simple instrumentation. Hence, the technology provides a platform that is potentially useful for clinical practice in pathology. It must be emphasized that no particular technology is ideally suited to all applications but a careful matching of the

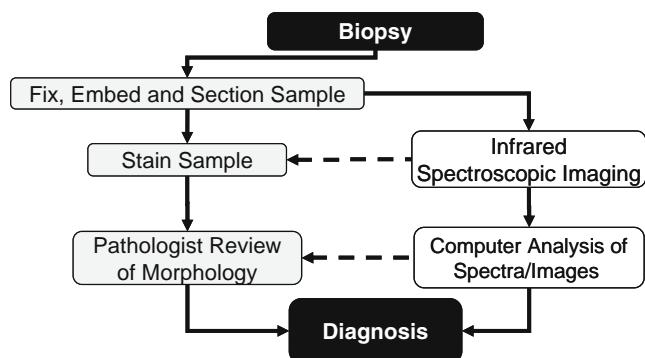


Fig. 2 Potential application of FTIR imaging for pathology. The current paradigm of cancer diagnosis and grading upon biopsy involves sample processing, staining, and pathologist review (left, shaded boxes). To implement the paradigm of automated analysis (right, unshaded boxes), IR chemical imaging is followed by computer analysis for diagnosis. Since IR imaging is label-free and non-perturbing, the sample can be stained, providing the pathologist with both IR chemical and conventional stained images

technique to the application can lead to useful protocols. While the potential advantages of FTIR imaging for examining tissue biopsies is high, practical protocols for clinical deployment are being developed by many groups.

Numerous recent reviews are available to address biomedical applications of FTIR spectroscopy and imaging [16–20], especially related to diseases and cancer. These reviews address instrumentation, the applicability to various

systems, spectroscopic bases and classification algorithms for decision making, and controversial aspects in the backdrop of the evolution of the field. The commercial availability of high-fidelity FTIR imaging instruments, advances in computers and data analysis algorithms, and increasing interest have combined to generate an increasing volume of studies. At the same time, there is considerable debate emerging on various aspects of the process. Reports study a variety of organs that may not correlate in behavior, utilize different sample acquisition and processing techniques, employ different instrumentation, data acquisition, or handling protocols, and apply a variety of decision-making algorithms. While this has led to a lively community of practitioners and exploration of various facets such as resolution, biological diversity, and chemometric or statistical methods, studies have generally focused on one aspect. Many excellent studies have developed each of these aspects to the point of routine use in advanced laboratory. The focus in the field is now on understanding biochemical signals and developing protocols from high quality data that can actually lead to clinical acceptance. We contend that the development of clinical protocols is necessarily integrative and, in this manuscript, review first the salient aspects in developing a practical, integrative approach to spectroscopic imaging for cancer histopathology. Second, we discuss the issues of spatial selectivity, sample size calculations,

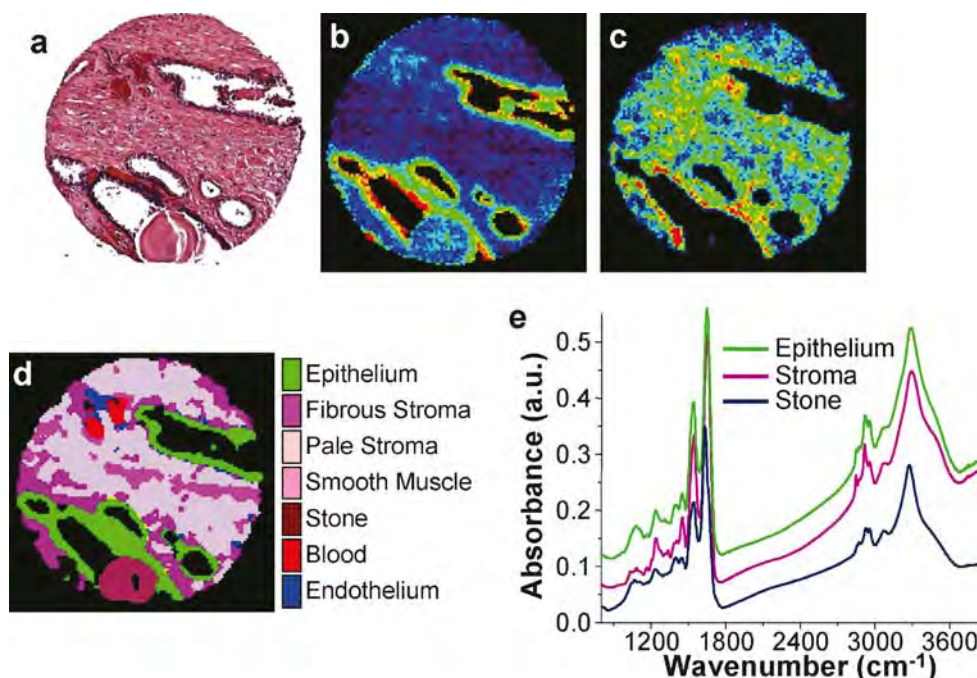


Fig. 3 Correspondence of conventionally stained and FTIR chemical images for pathology applications. **a** Hematoxylin and eosin (H&E)-stained image of prostate tissue section. Hematoxylin stains negatively charged nucleic acids (nuclei & ribosomes) blue, while eosin stains protein-rich regions pink. The diameter of the sample is ca. 500 μm . Simple univariate plots of specific vibrational modes provides for enhancement or suppression of specific cell types. **b** Absorption at

1,080 cm^{-1} commonly attributed to nucleic acids, highlights nucleic acid-rich epithelial cells in the manner of hematoxylin. **c** Spatial distribution of a protein-specific peak (ca. 1,245 cm^{-1}) highlights protein-rich regions in the manner of eosin. The entire spectrum can be analyzed for a series of markers that provide more information than H&E or univariate images, as shown in **d** where specific cells are color coded based on their spectral features (**e**)

optimization considerations, and potential improvements in algorithms that can provide faster results. Tests to determine performance and limits of accuracy are reported as a function of experimental parameters. We focus here on prostate histology as an illustrative test case, but emphasize that the approach is applicable and similar insight is gained with other tissues [21]. Further, exciting results have recently been reported for diagnosis, grading, and classification of prostate cancer [22–26], including the effects of zonal anatomy [27] and cytokinetic activity on spectra [28]. An extension of the methodology here to pathology will help formulate better protocols and allow a better understanding of the performance of classifiers.

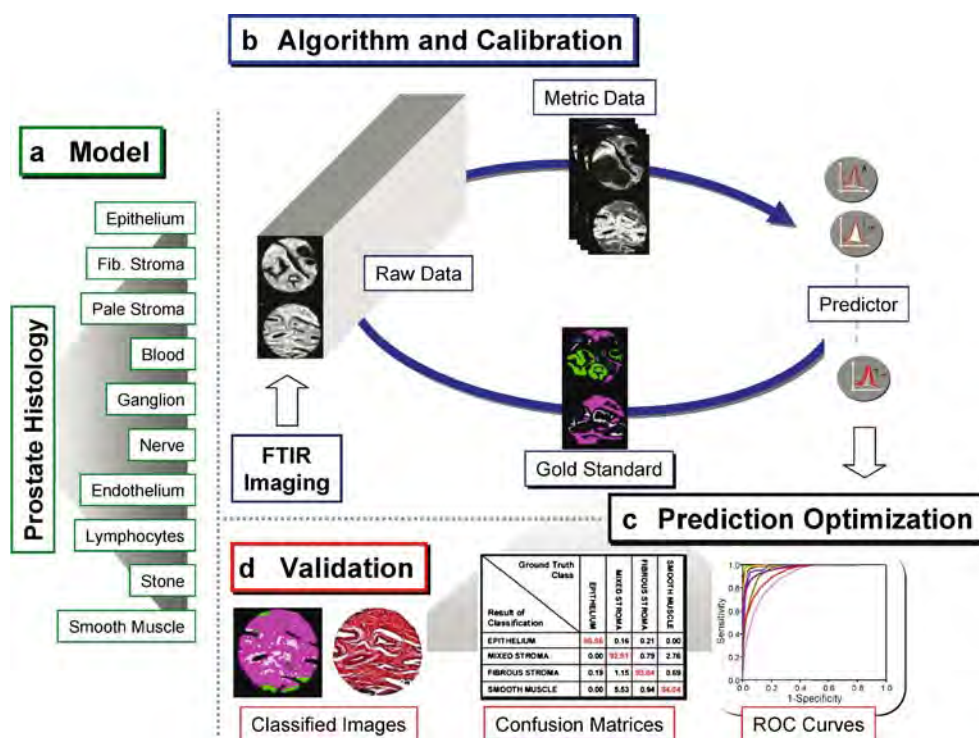
Approach and essentials

The promise of chemical imaging for pathology is illustrated in Fig. 2. Our approach has been to attempt integration of our developments with current clinical practice. Hence, we employ tissues that have been biopsied, fixed, embedded, and sectioned as per usual clinical protocols. We differ in the de-paraffinization step, suggesting a gentle wash with hexane and do not stain the tissue. Additionally, as IR chemical imaging only employs benign light, it is non-perturbing and entirely compatible with all downstream pathology processes. Hence, the sample may be stained as usual (Fig. 2, dashed arrow, top). Visualizations similar to those observed in conventional pathology are possible without staining the tissue. For example,

Fig. 3 correlates H&E and infrared spectral images. Visualizations similar to H&E images may be “dialed-in” by utilizing specific spectral features indicative of tissue chemistry. Although, the IR data only demonstrate univariate representations in the images, automated mathematical algorithms can determine the cell types and their locations within the image, while providing quantitative measures of accuracy and statistical confidence in results [29]. These data may be employed to directly provide diagnoses or to inform the pathologist (Fig. 2, dashed arrow, bottom), helping them make better decisions. Since the results are images, information exchange between spectroscopists and clinicians is facilitated. Spectroscopic analyses can potentially be fully automated; thus, no additional users need to be trained or knowledge base acquired by current clinicians.

A major challenge in the field is the development of robust algorithms that employ spectral data to provide histopathologic information. Both supervised and unsupervised approaches have been employed. We believe that unsupervised methods are more suited to research and discovery. Supervised methods are preferred when the data need to be related to known conditions, e.g., clinical diagnoses. The development of supervised classification of IR chemical imaging data for histopathology is fairly straightforward [30]. The process is shown in Fig. 4. First, a model for classification is selected. The model comprises all possible outcomes for any pixel in the images and is, hence, bounded by definition. We term each histologic constituent of the model a class to denote that it may not correspond to specific cell types or entities corresponding

Fig. 4 Process for relating pathologic or physiologic state to FTIR chemical imaging data. A model is chosen for supervised classification (**a**). **b–d** Training data is reduced in size and optimized into a prediction algorithm using gold standard data. The developed algorithm is validated against a second, independent data set and the accuracy is measured using three different methods: ROC curves, confusion matrices, and image comparisons



to morphology-based pathology. While this allows for simplifications and allows the user to focus on specific cells relevant in disease, it is also likely to prove useful in the discovery of different chemical entities that appear morphologically identical.

Next, data from a large number of tissue samples is recorded. A set of pixels are specifically marked (gold standard) by different colors to correspond to known regions of tissue, usually by comparison with an H&E-stained image or with immunohistochemically stained images [21]. The recorded data set is reduced to a smaller set of measures that capture the classification capability of the entire data set. We termed these measures metrics. There are numerous means of obtaining the metric data set: manual selection of large spectral regions, principal components analysis, genetic algorithms, or a sequential forward selection algorithm. A numerical algorithm is then chosen, for example, a linear discriminant analysis, neural network, SIMCA, or modified Bayesian classifier [31]. The classifier is optimized iteratively, if needed, to optimally predict the training data set. Subsequently, the algorithm is applied to a second data set (independent validation) that has been independently marked for each class. A comparison of the gold standard marking with the computationally predicted class provides a measure of the accuracy. We have employed three measures of accuracy: receiver operating characteristic (ROC) curves [32] that represent the sensitivity and specificity trade-off of the classifier, confusion matrices that provide the fraction of pixels of each class classified as pixels of all classes, and classified images that can be compared pixel-for-pixel to other images. Additionally, it is often instructive to drill into the classifier to obtain the basis for classification or the distribution of confidence intervals for various samples. The last two factors are generally not apparent in previous studies.

There are three key developments that are needed for this approach to be successful: (a) high-fidelity FTIR imaging instrumentation, (b) high-throughput sampling, and (c) robust classification that provides statistically significant results in a manner that can be appreciated by non-experts in spectroscopy. We briefly review the three developments next.

FTIR imaging

Need for spatially resolved data

The need for spatially resolved data has been recognized [33], but the effect of limited resolution data on classification is not entirely clear. The primary complication of coarse spatial resolution, obviously, arises from boundary

pixels. These can be defined as pixels that are assigned to one class but would likely yield more classes, to their physical limits, were finer resolution available. As a consequence, the spectral content of the boundary pixel is likely to be mixed and will likely lead to errors in classification. For example, the confounding contribution of stromal spectra to cancerous epithelial cells in breast tissue has been proposed [34]. As the resolution becomes coarser, the fraction of pixels in an image that belong to boundary pixels increases. Inclusion of these pixels has been shown to be a primary contributor to error rates in data [29], while their exclusion in accounting for accuracy necessarily implies that not all pixels are included. We sought to examine the effect of spatial resolution on the prevalence of boundary pixels.

We binned data acquired at 6.25- μm pixel size from 148 samples in a validation data set (≈ 7000 pixels/sample) to 10-, 15-, 20-, 30-, and 50- μm pixel sizes. There is an important distinction between pixel size and spatial resolution. The pixel size denotes the best possible optical resolution, which may be limited by longer wavelengths in the spectrum and optical effects to yield a poorer measured resolution [35–38]. For each dataset, we classified the tissue and determined neighbors of each pixel that did not belong to the class of the pixel. Some pixels that have no neighbors of other classes may still have empty pixels as neighbors. Since neighboring empty pixels can only provide optical distortion [39] but do not affect spectral content; we do not consider them further. The number of neighbors for epithelial pixels for different spatial resolutions may be seen in Fig. 5. The first observation is that a large majority of pixels have the same class pixels as all eight neighbors. The fraction of pixels with all neighbors of the same class

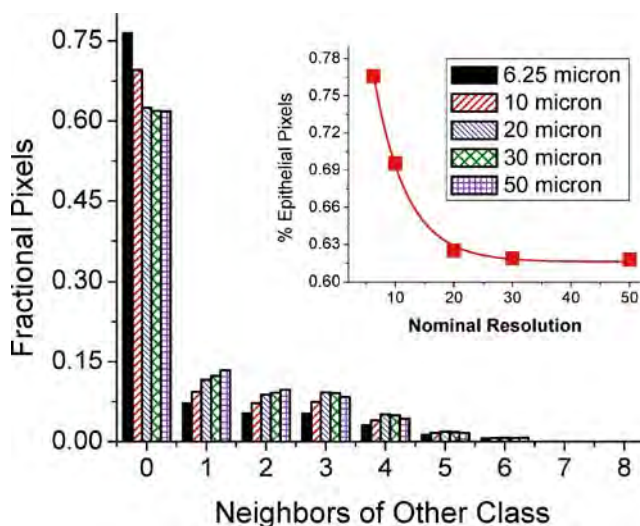


Fig. 5 Neighbors of cell types other than epithelium or empty space for different spatial resolutions. The inset shows the decrease in percent epithelial pixels that do not have any other cell types as neighbors

decreases rapidly with decreasing resolution and stabilizes at ca. 20 μm . Hence, a spatial resolution coarser than 20 μm is unlikely to have an effect on the classification but is expected to lead to about 25% more epithelial pixels being contaminated compared to 6.25- μm pixel sizes. The precise effect on a specific sample is very dependent on the sample morphology and is generally associated weakly with pathologic state. While in itself, the statistic does not imply that results from coarser resolution studies will be invalid, practitioners must recognize that error rates may be higher and that this contribution may be mitigated by using commonly available imaging systems.

One danger of classifying mixed composition pixels is whether they may be classified as an entirely different class or disregarded from the data set as belonging to no class. We simulated pixels of composition ranging from 0 to 100% for pairs of each class. We also added noise to simulate different data acquisition conditions. An example of the data can be seen in Fig. 6. Average spectra, one each from the two classes, are baselined and added in ratios varying linearly from 0 to 100%. Figure 6b demonstrates the classification of the gradient data set. In general, the classification works well, favoring the class with higher concentration. The classifier is also stable at the noise levels examined. A surprising result is that pixels between epithelium and fibroblast-rich stroma are classified as mixed stroma. This drawback, however, is the only example of two classes mixing to yield an entirely different one. The reason also stems from the definition of the mixed stroma class. While the class was designed to handle those

stromal cells that were not clearly fibroblasts or smooth muscle in origin but appeared mixed, a mix of epithelium and fibroblast-type stroma also leads to the classification as mixed stroma. Noise seems to have little effect on this behavior.

The full simulation of all classes (not shown) reveals that mixed pixels generally can be classified as the constituent classes with the higher concentration. Clearly, boundary pixels at epithelial fibroblast-rich regions must be handled with care. The increase in boundary pixels at lower resolution also implies that this type of systematic mis-assignment may arise more frequently. The rate of occurrence of boundary pixels may be even lower for synchrotron-based imaging that is conducted at higher pixel density or in emerging approaches that utilize synchrotron-based interferometers and array detectors. The simulated example above, however, demonstrates that simply over-sampling a spatial region to increase pixel density may allow for better definition of the interface and assignment of pixels, though it will not address spectral purity. Hence, for analyses based on spectral discrimination, mixture models will have to be developed based on entire spectra. For example, multivariate curve resolution techniques hold promise.

A further complication arises in using data from histologic classification for pathologic diagnoses. For example, the boundary epithelial pixels classified above may disproportionately contribute to classification errors. We have found evidence for the same in studies for both cancer pathology and for histology in tissue from different organs. For example, the boundary pixels in benign tissue get

Fig. 6 Mixture models and classification for prostate histology. **a** Absorbance at 1,080 cm^{-1} for three classes and their mixtures. The *first column* contains mixtures of epithelial cell spectra with the average spectrum from fibroblast-rich stroma and mixed stroma. The *second and third columns* contain mixtures with fibroblast-rich and mixed stroma, respectively. The concentration changes from 0 to 100% linearly along the *y-direction* as indicated by the *color bar* in **c**. **b** Along the *x-axis* of the composite image, the noise in each cell increases linearly. *Error bars* are standard deviations of noise in the spectra. **c** Classified image for the data, demonstrating the effect of composition and noise on classification. **d** Probability profiles of the three cell types at columns 1 and 25, demonstrating the effect of noise

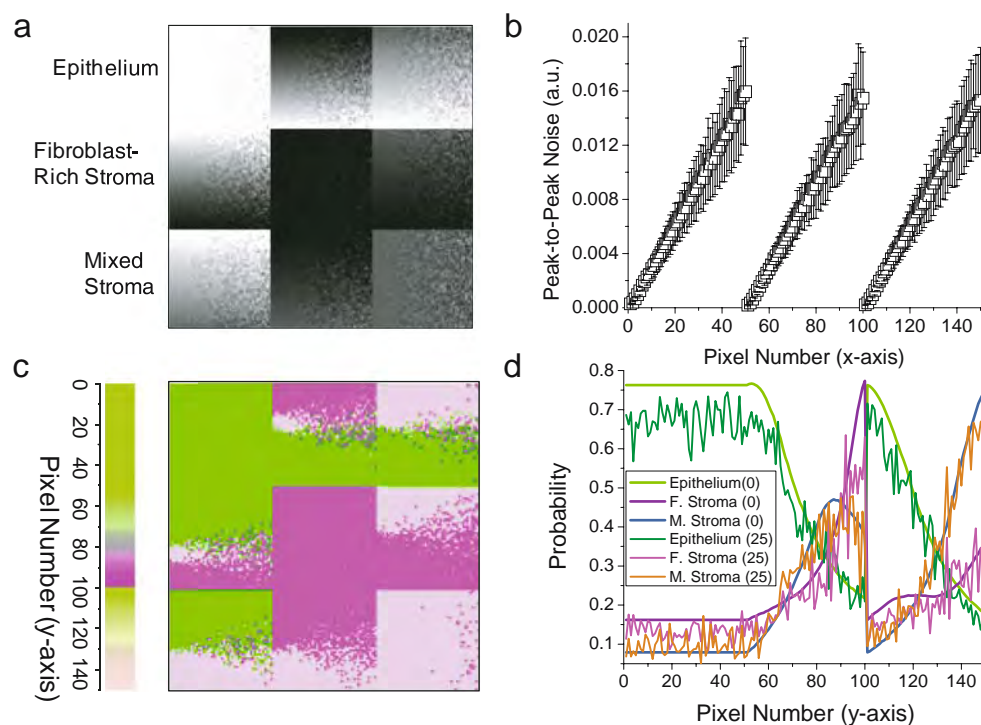


Table 1 Correlation of composition for samples between 6.25- μm pixel sizes and other pixel sizes

Pixel size (micron)	Epithelium	Fibroblast-rich stroma	Mixed stroma
10	0.9913x(0.9976)	0.9847x(0.9923)	1.0300x(0.9957)
20	1.0156x(0.9906)	0.9671x(0.9775)	1.0473x(0.9787)
25	1.0404x(0.9896)	0.9768x(0.9624)	1.0262x(0.9617)
30	1.0720x(0.9773)	0.9683x(0.9507)	1.0175x(0.9363)
50	1.1180x(0.9459)	0.9410x(0.8947)	1.0390x(0.8723)

The first row in each cell denotes the composition factor for that pixel size and class. For example, for every 100 μm^2 , the area of epithelial pixels at 10- μm pixel size is 99.13% of that at 6.25- μm pixel size. Increasing/decreasing numbers represent pixels being increasingly/decreasingly classified as that class. The ratios are not uniform for every sample and the regression coefficient of the best fit line passing through the origin is provided in the second row of the each table cell. Increasing pixel sizes reflect greater variance from the fit line

misclassified as cancerous, leading to the major source of error in applying this approach to pathology. At this time, the evidence is anecdotal and needs further investigation to quantify the extent of the error and its mitigation by advanced numerical processing. The last interesting aspect of lower spatial resolution is that it tends to over-predict certain classes. For example, Table 1 demonstrates the regression results of each samples composition against that obtained at 6.25 μm for three classes. While the regression coefficient is high, it is clear that epithelial and mixed stroma fractions are overestimated and fibroblast-rich stroma is underestimated with decreasing pixel size. There are differences based on underlying pathology. For example, normal epithelium is generally encountered in 10- to 40- μm -wide strips, while high grade tumor may be hundreds of micrometers to millimeters in size. Individual sample variability reflected in the regression coefficient decreases with increasing pixel size. In spectroscopic models to predict diseases that include morphological units but are based on average spectra, mixed pixels may lead to estimates with large errors. For example, a 1:1 mixed region of epithelial and fibroblast pixels at 6.25- μm pixel size increases to ca. 1.19:1 for 50- μm pixel size. Hence, the use of histologic mixture models at limited spatial resolution may not be estimated correctly, providing evidence that the percentage content of cell types in a limited field of view is likely to be a less robust measure of tissue histopathology.

Evolution and capabilities of current instrumentation

To overcome confounding by mixing, as discussed above, microspectroscopy was proposed as an alternative [40]. Single spectra (non-FTIR) have been recorded from microscopic samples for over 50 years [41] by restricting light incident on the sample through an aperture. More than one point, however, is required for tissue analysis. Hence, sequentially rastering the point at which spectra are recorded, termed mapping or point microscopy, was proposed [42]. A practical instrument obtained by coupling an interferometer, a microscope, and automated stage in the

late 1980s [43] helped in numerous materials science [44], forensic [45], and biomedical [46, 47] studies. Unfortunately, the mapping approach has a number of drawbacks in realizing the goal of an FTIR microscopy analog to optical microscopy [48].

More than 85% of cancer arises in epithelial cells, which often form surface layers that are 10- to 100- μm wide. As we demonstrated in the previous section, however, a resolution higher than ca. 10 \times 10 μm is preferable. Consequently, the illuminated spot at the sample has to be made smaller, throughput decreases proportionally, which in turn decreases the signal to noise ratio (SNR) of acquired spectra. Orders of magnitude brighter sources, e.g., synchrotrons, may be employed to recover the lost SNR. Unfortunately, synchrotron or free electron lasers [49] are prohibitively expensive and no laboratory lasers exist for the wide spectral region. An alternative is to average successive measurements (co-adding) to increase statistically the SNR. Since the SNR increases only as the square root of the number of averaged spectra, long averaging periods are required. The situation may be mitigated by using higher condensing optics, sources at higher temperatures, slightly faster scanning than used here,¹ gain ranging [50], or ultra-sensitive detectors [51]. Even if a hypothetical instrument with all these advances were constructed, ca. 10- to 20-fold reduction in time would be obtained. Furthermore, this calculation underestimates the time required by not considering losses due to diffraction or stage movement.

In prostate tissue, for example, the situation is similar to Fig. 1. Epithelial cells form 10- to 35- μm -wide foci around the cross-sections of ducts. Ducts appear as white circles in Fig. 1b, surrounded by epithelial cells that are depicted in blue. To analyze this morphology, aperture dimensions of ca. 6 $\mu\text{m}\times$ 6 μm (\approx cell size) are proposed [31]; for this case, the mapping approach would require ca. 1,028 h for a

¹ There is no advantage to faster scanning once the modulation frequency has reached optimum level for MCT detectors (1 MHz). The reduced time to observe signal then decreases the SNR.

500 $\mu\text{m} \times 500 \mu\text{m}$ sample [31]. Hence, mapping is not a viable option. In contrast to point mapping using apertures, large fields of view are measured in FTIR imaging. Contributions from different sample areas in imaging are separated by an array of mid-IR-sensitive detection elements in the manner of imaging with CCD devices for optical microscopy. By coupling the multichannel detection of focal plane array (FPA) detectors with the spectral multiplexing advantage of interferometry, an entire sample field of view is spectroscopically imaged in a single interferometer scan [52]. Depending on the microscopy configuration, thousands of moderate resolution spectra can be acquired at near-diffraction-limited spatial resolution in minutes [53, 54]. The time advantage over mapping is nominally the number of pixels in the FPA (16- to 65,000-fold) but the noise characteristics of FPAs are poorer than sensitive single point detectors [55]. Hence, the SNR-normalized advantage is lower [56]. Faster detectors are being used for imaging and promise significantly higher SNR in the same time. For example, we have employed a 128×128 element MCT array operating at ca. 16 kHz to acquire a full data set in ca. 0.07 s [unpublished]. These rates of data acquisition are approximately a factor of 10 higher than commercially available, but are required for practical data acquisition times. Increase in data acquisition speed remains a bottleneck for applications of IR imaging to routine clinical studies. Coupled with the complexity and cost of instrumentation, present technology provides preliminary capability but is likely to prove a barrier to practical clinical translation.

High-throughput sampling and statistical pitfalls

Quantitative analyses of results

The best imaging instruments (which employ sensitive detectors and a small multichannel advantage) can acquire data in about 0.1% of the time required for mapping for equivalent parameters. Hence, point mapping studies in pathology typically exceed numbers in only one of these categories: spatial resolution (ca. 15–20 μm), numbers of patients (ca. 50) or recorded small numbers of spectra per patient (ca. 100). These numbers may typically be improved an order of magnitude with imaging. For example, a recent report analyzed ca. ten million spectra from ca. 1,000 samples at a spatial resolution of 6.25 μm [26]. This quantitative validation is necessary for any automated biomarker approach (vide infra) [57]. Studies are underway in our and other laboratories to correlate spectral patterns with other physiologic and pathologic conditions; recent published studies verify the robustness and potentially wide applicability of FTIR microscopy [58, 59].

Sample size

Though these studies demonstrate potential, [60, 61] considerable debate exists on reproducibility and accuracy measures for larger studies [29]. The first response of many practitioners to new data is a question of validity based in limited statistical confidence. A detailed understanding is emerging from the work of several groups regarding appropriate sample control [62] and confounding factors due to biology [63]. Inherent differences between patient cohorts, effects of sample preparations and measurement noise are topics that can be addressed with the available imaging technology but are yet to be fully explored. Hence, validating robust spectral markers for large sample populations [64, 65] is exceptionally challenging and the chance for chance and bias influencing results exists.

Most importantly, the fundamental question of sample size required has remained open. There are two major concerns: first, the optimal sample size in forming calibration sets and a prediction algorithm. Second, investigators must determine whether the results shown can be supported by statistical considerations. While the first problem is essentially one of optimizing a model and prediction algorithm, the second impacts the quality of results and claims of applicability directly. In this manuscript, we examine only the second aspect. Determining the optimal sample size to form robust models is a more involved problem and is discussed elsewhere. The statistical validity of obtained results and dependence on data acquisition parameters are discussed later in this manuscript. Specifically, we estimate sample size based on the standard error for the area under the curve for an ROC curve.

Gold standard

The selection of pixels as gold standards needs great care. It must be done independently of any classifier training or validation, thus ensuring a blinded study design. Once the gold standard set is determined, it must not be changed. This will ensure that there is no bias in the process. Care must be taken to avoid pixels that do not lie on the tissue or those that are at the boundary as these may artificially inflate the error. The use of all pixels in an image has been suggested and their exclusion has been proposed to contribute selection bias. Selection bias, however, does not arise in pixels that are chosen independent of validation algorithms. The exclusion of boundary pixels is necessary in both training (to avoid spurious probability distribution functions) and validation (to prevent introduction of errors). There are major technological difficulties in relating stained visible to IR images from unstained tissue due to changes during staining, leading to errors. Hence, it has been proposed that the exclusion of boundary pixels in akin to

the performance of a classifier with a reject option for the boundary.

Sampling, archiving, and consistency

While it is unclear what an optimal sample size would be, it is clear that a large number of tissue samples are needed for effective validation. While it may theoretically be possible to train on a single sample, validation of a protocol is required on more samples. We recognized that one does not need to observe the full surgically resected tumor for validating IR protocols, but would need a representative small section. Hence, we employed tissue microarrays (TMAs) [66] as a platform for high-throughput sampling. TMAs consist of a large number of small tissue samples arranged in a grid and deposited on the same substrate. They are typically manufactured by embedding cylindrical cores in a receiving block and sectioning the block perpendicular to the long axis of the core. Thin sections are then floated on to a rigid substrate for analysis. The technique facilitates rapid visualization of results of any classification protocol, while revealing localization and prevalence of any errors. Sample processing times may easily be increased 100-fold, valuable tissues are optimally utilized, and consecutive TMA sections can be used to correlate with staining results. Construction and analysis of TMAs has been automated, further increasing the throughput. For spectroscopists, TMAs provide a ready source of tissue to test hypothesis and develop prediction models.

The validity of employing TMAs for prostate cancer research and, especially, for cancer grading has been addressed by a number of authors [67]. For example, a study of genitourinary pathologists [4] with images from TMA cores assesses that ca. 90% considered this approach useful for resident training and for pathology teaching. Further, Gleason score was easily assigned to each TMA spot of a 0.6-mm-diameter prostate cancer sample. Hence, the utility of TMAs is not only in providing numerous samples in a compact manner for the advantages above, but also in consistency of the diagnoses and precision in analyzing similar areas. Virtual tissue microarrays could be constructed from different areas of large samples, thus providing many sub-samples for within-patient and among-patient comparisons. This approach has not yet been reported but is likely a useful extension of the TMA concept.

Prediction algorithms and high-throughput data analysis

Univariate algorithms

The major technological advances of fast FTIR microscopy and high-throughput tissue sampling have been addressed

by imaging and TMAs respectively. There is still some confusion and widespread disagreement, however, about the “best” approach to extract histopathologic information from FTIR imaging data. Several early manuscripts employ univariate correlations to disease states [68]. While the results were exciting, it is now realized that they were statistically flawed and did not necessarily contain a fundamental basis in cancer biology. To our knowledge, there is no manuscript that has expressly demonstrated, using statistics arguments, why univariate analyses are likely to fail. There is widespread consensus and anecdotal evidence, however, among practitioners that argues against the approach. Consider the distributions for a univariate measure (absorbance at $1,080\text{ cm}^{-1}$ that is normalized to the amide I peak height) for benign and malignant cases as shown in Fig. 7.

The normalized histograms reveal that for specific, single samples the distribution of absorbance at pixels is such that it clearly indicates the metric to be a good one for cancer discrimination. When the distribution from all samples is considered, however, there is little difference in the distributions. Hence, many univariate measures described in the literature do not hold up in wide population testing. A TMA-based, high-throughput validation can easily prove that the measure is not a good one but does discriminate some samples. In Fig. 7, a cutoff value can generally be found that distinguishes disease, leading to the erroneous conclusion that the feature is universally indicative of disease state. Since a typical infrared spectrum has numerous frequencies and even non-chemically specific features that can provide discrimination, a small number of samples increases the probability of finding such discrimination by chance alone. Univariate measures that appar-

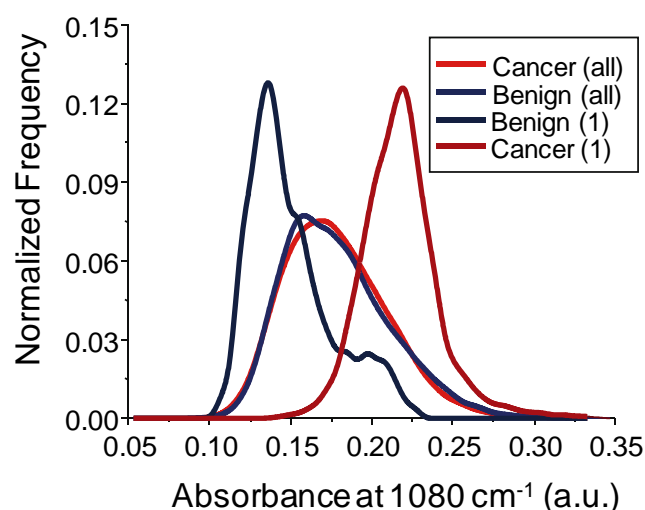


Fig. 7 Distribution of absorbance for individual spots and all pixels from each class, normalized by the total number of pixels in the class, demonstrates that the examination at patient level and at a global level may not correspond

ently provide discrimination when none exists can be equated to the false discovery rate (FDR) [69] of metrics. The FDR is very different from the p -value for determining that a metric separates two distributions; a much higher FDR can be tolerated than can a p -value. Similarly, a false negative rate has been proposed [70], which is not critical for our case as we have observed high accuracy without use of any erroneously left out metrics. While detailed calculations and their underlying concepts are too lengthy to reproduce here, for the sake of completeness, it suffices to say that for the expected number of metrics demonstrating discrimination, the FDR tends to zero for larger than ca. 30 samples. While correlations due to chance can be minimized by this approach, there is potential for unknown bias or error in prediction for small numbers of samples. Hence the algorithm must be integrated with sampling considerations.

Multivariate algorithms

It was argued in the previous section that univariate analysis may not provide a good measure of the population distribution. It can alternatively be argued that the individual differences in univariate measures are masked if population measures of the same are employed. Similarly, multivariate techniques may mask the individual measures in population testing. Hence, our philosophy has been to employ a multivariate, supervised classification in which the metrics are derived from univariate analyses. This enables us to carefully examine each metric for both population as well as individual sample relevance. While unsupervised clustering approaches provide good insight into spectral similarity, a supervised method forces a relation to common clinical knowledge. For example, as shown in Fig. 4 for prostate tissue, we consider a ten-class model to determine histology. The drawback is that the sensitivity of the approach to individual samples is lost at the expense of generality. One could potentially combine clustering and supervised classification. Clustering information on the training data set would emphasize individual sample distributions, which would allow for supervised classification tailored to each cluster type. Such an approach has not been implemented yet but is being attempted in our laboratories to classify samples optimally.

Dimensionality reduction

It is well recognized that the spectrum at each pixel needs to be reduced to a smaller set of useful descriptors that capture the essential information inherent in the spectrum. The reduction of full spectral information to essential measures helps eliminate from consideration those spectral features that have no information (non-absorbing frequen-

cies), little biochemical significance (e.g., apparent absorption at non-chemically specific frequencies), inconsistent measures that may degrade classification, and those with redundant information. The number of useful measures is significantly smaller than the number spectral resolution elements and, hence, the process is also termed dimensionality reduction. Dimensionality reduction and further refinement (vide infra) also helps reduce the incidence of prediction by chance alone, reduce computation time and storage requirements. Potential measures of a spectrum's useful features are termed metrics and are defined manually in our scheme.

It may be argued that the metrics are not selected in an objective manner due to a human performing this task and some computer routines must be employed. While the use of an automated computer program is most certainly objective and reproducible, the algorithm that drives such programs is generated from spectroscopy knowledge. A well-trained spectroscopist can recognize spectral features and assign them to appropriate their biochemical basis. While a computer algorithm may be able to enhance subtle features in the spectrum, automated peak-picking algorithms run the risk of substantial error as they are based on some very specific criteria that may not be universally valid. We believe that computer algorithms are more suited to finding correlations and patterns that a human cannot for the sheer size and complexity of data. Hence, the process of determining which spectral features to consider is entirely manual in our approach. It must be emphasized that the universal set of metrics is selected manually but that the data reduction step to a set of metrics to be used in algorithms is entirely based on objective algorithms. Manual refinement of metrics for classification is, obviously, not recommended for possibilities of overlooking specific features, biasing the selection to specific feature sets, or in determining the optimal set of metrics for a classifier. Dimensionality reduction is also intimately linked to the data quality and classification algorithm employed.

Classification algorithm

A number of supervised algorithms have been applied to dimensionally reduced data, including those based on linear discriminant analysis, neural networks, decision trees, and modified Bayesian Classifiers. An intermediate step in some of these algorithms provides for a fuzzy result in which every pixel has a probability of belonging to every class. For example, in our approach, each pixel can have a probability (between zero and one) of belonging to each class. A discriminant function then assigns each pixel to a class based on a decision rule. The pre-discriminant data set, termed rule imaging set, contains important information. In our algorithm, it is a direct measure of the

probability of the pixel belonging to the class. Hence, the probability value may be used to compare the potential of two protocols to distinguish a cell type or to quantify confidence in results for tissue classified by different methods.

Measures of accuracy and optimization

We prefer the use of the AUC for both optimizing algorithms and for validating results. Confidence in the value of the AUC, hence, is the primary test for the validity of developed algorithms and is characterized by the standard error of the value. For example, in validating the discrimination of epithelial from stromal pixels in a blinded validation set, the cumulative distribution of AUC in a TMA is shown in Fig. 8. More than 20% of the spots had an AUC >95% and no AUC value below 0.8 was recorded. One drawback of using ROC curves and AUC values is that the results are valid for one at a time classification. Hence, we have analyzed here the segmentation of epithelium from all other cell types. The tissue is classified into ten classes as before but the results are lumped into epithelial and non-epithelial pixels. Further, not all TMA cores have all types of cells. Hence, the two-class model also allows us to examine a large number of samples. Last, we excluded cores that did not contain at least 100 pixels of each class to leave 103 cores for the analysis.

Quantitative measures of performance and accuracy are perhaps the weakest portion of reports using IR spectroscopy for cancer pathology. Typically, sensitivity and specificity have been employed as summary measures. While these are indeed very relevant, we demonstrate that they are insufficient and classification analysis must utilize more measures to understand the process. Specifically, the use of

receiver operating characteristic (ROC) curves [71] is an excellent direction. The area under the ROC curve is a further summary measure that provides both a quantitative understanding of the discrimination potential of the model and a convenient measure to compare multiple classification models. The third tool we introduced was the confusion matrix. While ROC curves provide the potential for correct classification of a binary rule at a time, confusion matrices correspond to a particular point on the ROC curve under the constraints of accuracy measures of other classes. These also directly correspond to the final segmentation of the rule image under an optimization condition. The optimization condition may simply be the maximization of the accuracy or may be the minimization of certain types of errors.

Discriminant and class assignment

In a multi-class analysis, our approach to evaluating ROC curves for a class is one at a time, i.e., all other classes are essentially lumped in the rule data and the highest probability of the lumped ensemble is compared to the class whose ROC curve is being built. Hence, the AUC values must be regarded as a potential for classification. They are best suited to answer the binary question of whether a pixel is correctly identified or not when considering a single class. This method is ideally suited to a cascaded classifier one at a time. Such a classifier has not been reported yet but would provide a means to explicitly determine the error for any given classification scheme.

Experimental parameters and classification

Here, we take advantage of the trading rules of FTIR spectroscopy and imaging to model the effects of the experimental parameters on the classification process. While the signal to noise ratio (SNR) and resolution are generally arbitrarily fixed in most studies, we demonstrate their importance in classification.

Effect of signal to noise ratio

There are two issues: what is the “best” SNR to formulate algorithms and second, provided an algorithm, what is the least SNR that would provide adequate classification. Only the latter issue is examined here. As with conventional FTIR spectrometers, imaging spectrometers obey the trading rules of IR spectroscopy. Hence, if an n -fold reduction in SNR provides the same results, data acquisition will be n^2 -fold faster. Thus, in addition to an interesting fundamental behavior of the classifier, the role of SNR has a direct bearing on the speed at which data is acquired.

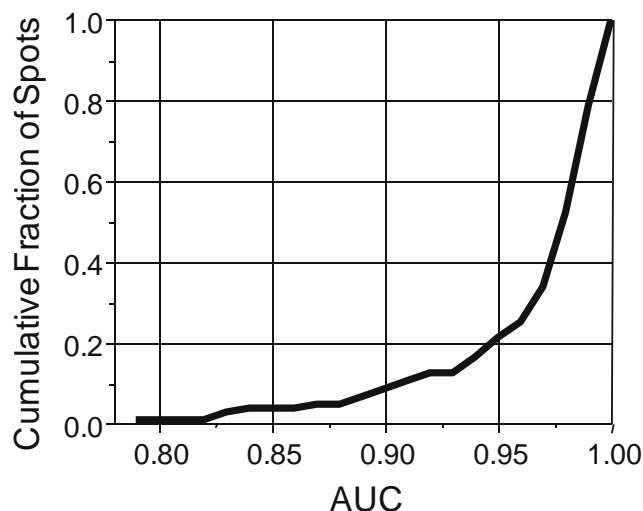
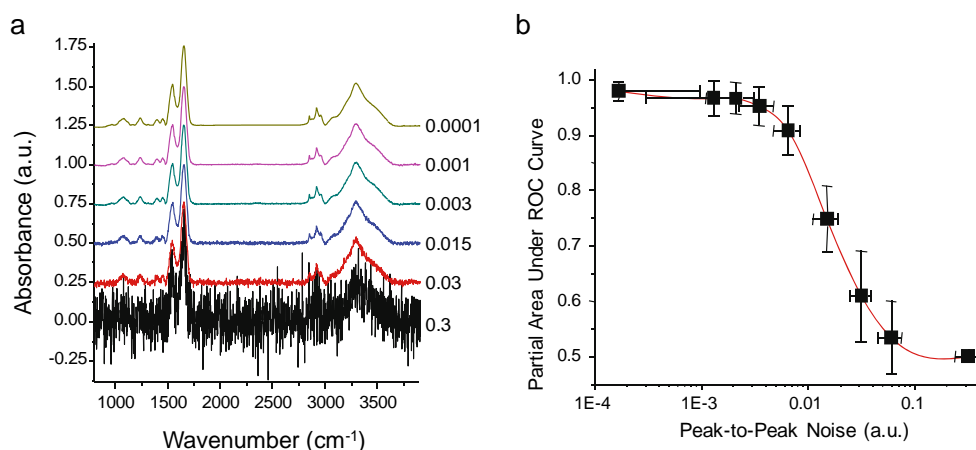


Fig. 8 Distribution of AUC values in a TMA for discriminating epithelium from stroma using the ten-class model

Fig. 9 **a** Noise in the data set as a function of added random noise. **b** Effect of spectral noise on the accuracy of classification as measured by AUC values



We examined classification accuracy as a function of average spectral noise. To strictly examine the effect of noise, data must be acquired at different co-added spectral numbers. The time required for imaging an array multiple times, however, is prohibitive. Hence, we computationally added random, Gaussian noise to the original spectral data. Peak-to-peak and root mean square (rms) noise were measured in the 1,950- to 2,150-cm⁻¹ region adjacent to the amide I peak.² Representative single pixel spectra from the data sets are shown, as a function of noise, in Fig. 9a. We additionally plotted the observed noise levels against the added noise to verify linearity (plot not shown). The linear relationship conforms to the expected result and provides a scaling factor to express the equivalent reduction in data acquisition time (co-addition) that would be realized at that noise level. For example, the addition of 0.005 a.u. of noise raises the peak-to-peak noise from 0.0013 to 0.015 a.u., corresponding to a decrease in data acquisition time by a factor of ca. 100 for this data set. In addition to increasing noise, we employed an algorithm based on an MNF transform [72, 73] to mathematically eliminate noise. The observed peak-to-peak noise was 0.00017 a.u., corresponding to an increase in data acquisition time by a factor greater than ca. 100. Hence, the data examined span about 5 orders in magnitude of collection time.

The average height of the amide I peak was 0.42 a.u. in all cases, providing a SNR of 2,500 (MNF-corrected data) to 1.5 for the data sets. Accuracy as a function of the noise level is shown in Fig. 9b. While the x-error bars indicate the standard deviation of noise levels in pixels, the y-error bars indicate the standard deviation in AUC values of all ten

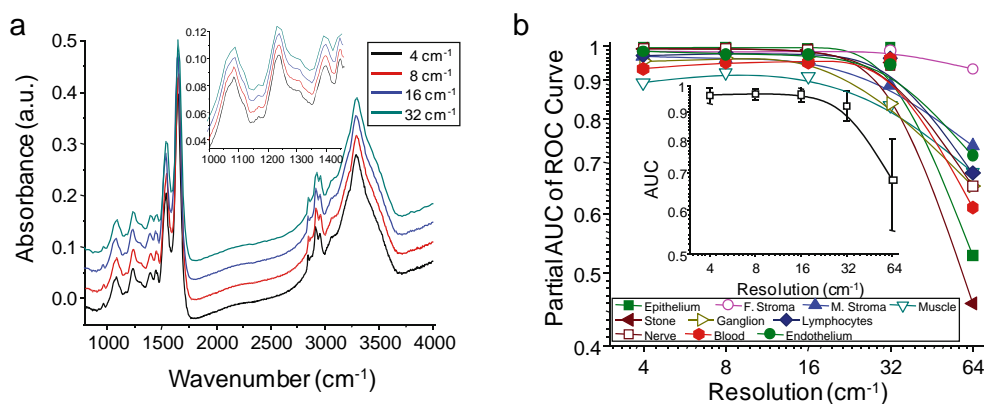
classes. As a general rule, the classification improves with lower noise levels. We first note that the classification does not become perfect for any noise level and there is a significantly diminishing return in increasing the SNR beyond a level. At the other end, the ability to distinguish classes is entirely lost at levels of ca. 0.1. Performance across multiple data sets observed using our prediction model indicates that the increases demonstrated at noise levels lower than ca. 0.003 a.u. are within the variance. Hence, there is little benefit to decreasing the noise levels below ca. 0.003 a.u. for this data set, or to increasing the SNR beyond ca. 150. It must be emphasized that the model, prediction algorithm, and discriminant function are intimately linked in a non-linear manner. While this makes it impossible to predict the behavior generally of all classification approaches, this simple exercise may be conducted to determine the optimal data acquisition parameters. For our selected metrics and model, it appears that the data acquisition time can be decreased by a factor of ca. 3 without significant degradation in accuracy.

Spectral resolution

We next examined the effect of spectral resolution on the results that would be obtained using the developed algorithm. As in the previous section, the data were not re-acquired but were downsampled from acquired data using a neighbor binning procedure. Spectra from the same epithelial class pixel, at different resolutions (Fig. 10a), demonstrate the effect of downsampling on feature definition. Figure 10b demonstrates, first, that the peak-to-peak noise levels over the region remain the same with spectral resolution. As previously observed, noise is an important control in comparing spectra; the peak-to-peak noise over the same number of data points was preserved by neighbor binning. In practice, the constant-throughput spectrometer would provide a SNR (or noise level, in this case) that decreases linearly with resolution. Since most array

² It is noteworthy that we are examining trends in the absorbance spectra. Strictly, SNR should be measured in single beam spectra to relate rigorously to theory. It can be shown, however, that the trend will hold approximately for the absorbance spectra as well. Many practitioners advocate the use of rms SNR. We are employing peak-to-peak fluctuations over the same spectral range. Hence, the noise values we obtain will be higher but will follow the same trend.

Fig. 10 **a** Spectra obtained by downsampling acquired data to different resolutions using a neighbor binning procedure. The *inset* demonstrates the effect of resolution on narrower features in the spectrum. **b** AUC values for each class and average AUC values as a function of spectral resolution demonstrate a decrease only for coarse spectral resolution



detectors can be operated with higher integration times, it is fair to assume that the time advantage in decreasing resolution would be linear. Second, the performance of the classifier is very nearly the same for finer spectral resolutions and degrades only significantly for 32 cm^{-1} . While the results may appear to be surprising, a closer analysis of the basis of the algorithms provides insight into the trends.

The classifier is based on absorbance and center of gravity measures of the peaks. It is well established that absorbance is measured accurately, provided that the FWHH of the peak is not significantly smaller than the resolution. The Ramsay resolution parameter, σ , is a useful measure that was originally developed for monochromators but has been shown to be applicable to FTIR spectrometers as well [74]. While most bands are broad and peak absorbance lower than ca. 0.7, absorbance values are not expected to be adversely impacted from the measurement process. With decreasing resolution, however, broadening within complex peaks shapes may lead to observed changes in the apparent absorption at a specific wavenumber. The change itself may not have a significant influence on the classifier performance as it depends on several such metrics. A second type of metric calculates the area under the curve. This is not expected to be impacted significantly for most peaks. The third type of metric we have used is the center of gravity of a spectral region. While spectral analyses ordinarily attempt to locate the peak position and use it as a metric, we chose the center of gravity for its sensitivity to both position and asymmetrical shape changes in complex spectral envelopes observed in biological samples. Since the classifier is based on center of gravity of a feature and not on the wavenumber of the peak maximum, it is a very robust measure that is relatively unaffected by spectral resolution or noise.

Generalization of developed algorithms to instruments and practical approaches

The characterization of classification with regard to spectrometer performance (SNR) and spectral resolution

provides information to optimize parameters on one spectrometer. It is unclear, however, if the calibration would transfer to another spectrometer. We contend that the potential for a successful transfer is high as the classification process is relatively insensitive to resolution, implying that it would only be weakly sensitive to apodization or to small inaccuracies in wavelength scale. Similarly, if the SNR of acquired data is used as control, perturbations due to fixed pattern noise in focal plane array detectors or the different use of electronic filters by different manufacturers is likely to be insignificant in classifying tissue correctly. Various instrument manufacturers also set the nominal optical resolution differently in their instruments. The issue of spatial resolution, of course, is more complex. Nevertheless, any resolution setting around the wavelength-limited case will likely provide consistent results. To our knowledge, there has been no comparison yet of classifier performance across mid-IR FTIR imaging spectrometers using algorithms developed on one specific instrument. The developed protocol provides for such a framework and detailed results are awaited from on-going work [75].

Outlook and prospects

An exciting period in imaging tissues spectroscopically with low power, optical microscopy-comparable resolution is emerging. Considerable work, however, needs to be accomplished before this idea can become a clinical reality. An ultimate goal of such studies is to provide a key technology for emerging molecular pathology. The approach promises greatly reduced error rates, automation, and economic benefits in current pathology practice. Looking to the future, chemical imaging approaches will be employed for diagnosing cancers in pre-malignant stages prior to their apparent changes observable by conventional means, predicting the prognosis of the lesion and intra-operative imaging in real-time. Fundamental studies in drug discovery and mechanisms of molecular interactions are further examples that would be enabled by progress in this

area. Doubtless, exciting applications lie ahead and progress is rapidly being made towards practical applications but much work needs to be done to carefully apply this powerful technology to multiple aspects of pathology. Success in this endeavor promises to change the practice of pathology radically and alter the clinical management of cancer patients.

Acknowledgement The author would like to acknowledge collaborators over the years, especially Dr. Stephen M. Hewitt and Dr. Ira W. Levin of the National Institutes of Health, for numerous useful discussions and guidance. Discussions and help from Dr. Daniel Fernandez during the formative years of this work are also appreciated. Funding for this work was provided in part by University of Illinois Research Board and by the Department of Defense Prostate Cancer Research Program. This work was also funded in part by the National Center for Supercomputing Applications and the University of Illinois, under the auspices of the NCSA/UIUC faculty fellows program.

References

1. Woolf SH (1995) *N Engl J Med* 333:1401–1405
2. Humphrey PA (2003) Prostate pathology. American Society for Clinical Pathology, Chicago
3. Partin AW, Mangold LA, Lamm DM, Walsh PC, Epstein JI, Pearson JD (2001) *Urology* 58:843–848
4. De La Taille A, Viellefond A, Berger N, Boucher E, De Fromont M, Fondimare A, Molinié V, Piron D, Sibony M, Staroz F, Triller M, Peltier E, Thiounn N, Rubin MA (2003) *Hum Pathol* 34:444–449
5. Levin IW, Bhargava R (2005) *Annu Rev Phys Chem* 56:429–474
6. Navratil M, Mabbott GA, Arriaga EA (2006) *Anal Chem* 78:4005–4019
7. Caprioli RM, Farmer TB, Gile J (1997) *Anal Chem* 69:4751–4760
8. Chaurand P, Schwartz SA, Billheimer D, Xu BGJ, Crecelius A, Caprioli RM (2004) *Anal Chem* 76:1145–1155
9. Kurhanewicz J, Vigneron DB, Hricak H, Narayan P, Carroll P, Nelson S (1996) *Radiology* 198:795–805
10. Lewis EN, Gorbach AM, Marcott C, Levin IW (1996) *Appl Spec* 50:263–269
11. Diem M, Romeo M, Boydston-White S, Miljkovic M, Matthaues C (2004) *Analyst* 129:880–885
12. Mendelsohn R, Paschalis EP, Boskey AL (1999) *J Biomed Opt* 4:14–21
13. Kidder LH, Kalasinsky VF, Luke JL, Levin IW, Lewis EN (1997) *Nat Medicine* 3:235–237
14. Ellis DI, Goodacre R (2006) *Analyst* 131:875–885
15. Bhargava R, Levin IW (eds) (2005) *Spectrochemical analysis using infrared multichannel detectors*. Blackwell, Oxford
16. Petrich W (2001) *Appl Spectrosc Rev* 36(2):181–237
17. Andrus PG (2006) *Tech Cancer Res Treat* 5:157–167
18. Krafft C, Sergo V (2006) *Spectroscopy* 20:195–218
19. Petitbois C, Deleris G (2006) *Trends Biotechnol* 24:455–462
20. Walsh MJ, German MJ, Singh M, Pollock HM, Hammiche A, Kyrgiou M, Stringfellow HF, Paraskevaidis E, Martin-Hirsh PL, Martin FL (2007) *Cancer Lett* 246:1–11
21. Keith FN, Bhargava R (2007) *Tech Cancer Res Treat* (submitted)
22. Gazi E, Dwyer J, Gardner P, Ghanbari-Siakhal A, Wade AP, Myan J, Lockyer NP, Vickerman JC, Clarke NW, Shanks JH, Hart C, Brown M (2003) *J Pathology* 201:99–108
23. Gazi E, Baker M, Dwyer J, Lockyer NP, Gardner P, Shanks JH, Reeve RS, Hart C, Clarke NW, Brown M (2006) *Eur Urol* 50:750–761
24. Harvey TJ, Henderson A, Gazi E, Clarke NW, Brown M, Faria EC, Snook RD, Gardner P (2007) *Analyst* 132:292–295
25. Paluszkiwicz C, Kwiatek WM, Banas A, Kisiel A, Marcelli A, Piccinini A (2007) *Vib Spectrosc* 43:237–242
26. Fernandez DC, Bhargava R, Hewitt SM, Levin IW (2005) *Nat Biotechnol* 23:469–474
27. German MJ, Hammiche A, Ragavan N, Tobin MJ, Cooper LJ, Matanhelia SS, Hindley AC, Nicholson CM, Fullwood NJ, Pollock HM, Martin FL (2006) *Biophys J* 90:3783–3795
28. Gazi E, Dwyer J, Lockyer NP, Miyan J, Gardner P, Hart CA, Brown MD, Clarke NW (2005) *Vib Spectrosc* 38:193–201
29. Bhargava R, Hewitt SM, Levin IW (2007) *Nat Biotechnol* 25:31–33
30. Srinivasan G, Bhargava R (2007) *Spectroscopy* 22:30–43
31. Bhargava R, Fernandez DC, Hewitt SM, Levin IW (2006) *Biochim Biophys Acta Biomembr* 1758:830–845
32. Swets JA (1988) *Science* 240:1285–1293
33. Lasch P, Naumann D (2006) *Biochim Biophys Acta* 1758:814–829
34. Jackson M, Choo LP, Watson PH, Halliday WC, Mantsch HH (1995) *Biochim Biophys Acta* 1270:1–6
35. Sommer AJ, Katon JE (1991) *Appl Spectrosc* 45:1633–1640
36. Carr GL (2001) *Rev Sci Instr* 72:1613–1619
37. Bhargava R, Wang SQ, Koenig JL (1998) *Appl Spectrosc* 52:323–328
38. Budevska BO (2000) *Vib Spectrosc* 24:37–45
39. Romeo M, Diem M (2005) *Vib Spectrosc* 38:129–132
40. Jackson M (2004) *Faraday Discuss* 126:1–18
41. Norris KP (1954) *J Sci Instr* 31:284–287
42. Rousch PB (ed) (1985) *The design, sample handling, and applications of infrared microscopes*. ASTM STP 949, American Society for Testing and Materials, Philadelphia
43. Kwiatkowski JM, Reffner JA (1987) *Nature* 328:837–838
44. Koenig JL (1999) *Spectroscopy of polymers*, 2nd edn. Elsevier, New York
45. Bartick EG, Tungol MW, Reffner JA (1994) *Anal Chim Acta* 288:35–42
46. Wetzel DA, LeVine SM (1999) *Science* 285:1224–1225
47. Gremlich H-U, Yan B (eds) (2000) *Infrared and Raman spectroscopy of biological materials (practical spectroscopy)*. Marcel Dekker, New York
48. Bhargava R, Wall BG, Koenig JL (2000) *Appl Spectrosc* 54:470–474
49. Vobornik D, Margaritondo G, Sanghera JS, Thielen P, Aggarwal ID, Ivanov B, Miller JK, Haglund R, Tolk NH, Congiu-Castellano A, Rizzo MA, Piston DW, Somma F, Baldacchini G, Bonfigli F, Marolo T, Flora F, Montemali RM, Faenov A, Pikuz T, Longo G, Mussi V, Generosi R, Luce M, Perfetti P, Cricenti A (2004) *Infrared Phys Tech* 45:409–416
50. Hirschfeld T (1979) *Appl Spectrosc* 33:525–527
51. Wetzel DL (2002) *Vib Spectrosc* 29:183–189
52. Carter MR, Bennett CL, Fields DJ, Hernandez J (1995) *Proc SPIE* 2480:380–386
53. Lewis EN, Treado PJ, Reeder RC, Story GM, Dowrey AE, Marcott C, Levin IW (1995) *Anal Chem* 67:3377–3381
54. Colarusso P, Kidder LH, Levin IW, Fraser JC, Arens JF, Lewis EN (1998) *Appl Spectrosc* 52:106A–120A
55. Snively CM, Koenig JL (1999) *Appl Spectrosc* 53:170–177
56. Bhargava R, Levin IW (2001) *Anal Chem* 73:5157–5167
57. Ransohoff DF (2004) *Nat Rev Cancer* 4:309–314
58. Bhargava R, Levin IW (eds) (2005) *Spectrochemical analysis using infrared multichannel detectors*. Blackwell, Oxford, pp 56–84
59. Various contributors (2006) *Biochim Biophys Acta Biomembr* 1758

60. Wood BR, Chiriboga L, Yee H, Quinn MA, McNaughton D, Diem M (2004) *Gynecol Oncol* 93:59–68
61. Malins DC, Polissar NL, Nishikida K, Holmes EH, Gardner HS, Gunselman SJ (1995) *Cancer* 75:503–517
62. Boydston-White S, Gopen T, Houser S, Bargonetti J, Diem M (1999) *Biospectroscopy* 5:219–227
63. Shaw RA, Guijon FB, Paraskevas V, Ying SL, Mantsch HH (1999) *Anal Quant Cytol* 21:292–302
64. Mansfield JR, McIntosh LM, Crowson AN, Mantsch, HH, Jackson, M (1999) *Appl Spectrosc* 53:1323–1333
65. McIntosh LM, Jackson M, Mantsch HH, Stranc MF, Pilavdzic D, Crowson AN (1999) *J Invest Dermatol* 112:951–956
66. Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schraml P, Leighton S, Torhorst J, Mihatsch MJ, Sauter G, Kallioniemi OP (1998) *Nat Med* 4:844–847
67. Camp RL, Charette LA, Rimm DL (2000) *Lab Invest* 80:1943–1949
68. Paluszkievicz C, Kwiatek WM, Banas A, Kisiel A, Marcelli A, Piccinini M (2007) *Vib Spectrosc* 43(1):237–242
69. Benjamini Y, Hochberg Y (1995) *J R Stat Soc Ser B* 57:289–300
70. Pawitan Y, Michiels S, Koschielnny S, Gusnanto A, Ploner A (2005) *Bioinformatics* 21:3017–3024
71. Stone N, Kendall C, Smith J, Crow P, Barr H (2004) *Faraday Diss* 126:141–157
72. Bhargava R, Wang SQ, Koenig JL (2000) *Appl Spectrosc* 54:486–495
73. Bhargava R, Wang SQ, Koenig JL (2000) *Appl Spectrosc* 54:1690–1706
74. Anderson RJ, Griffiths PR (1975) *Anal Chem* 47:2339–2347
75. Llorca X, Reddy RK, Bhargava R (in preparation)

Observer-invariant histopathology using genetics-based machine learning

Xavier Llorà · Anusha Priya · Rohit Bhargava

© Springer Science+Business Media B.V. 2007

Abstract Prostate cancer accounts for one-third of noncutaneous cancers diagnosed in US men and is a leading cause of cancer-related death. Advances in Fourier transform infrared spectroscopic imaging now provide very large data sets describing both the structural and local chemical properties of cells within prostate tissue. Uniting spectroscopic imaging data and computer-aided diagnoses (CADx), our long term goal is to provide a new approach to pathology by automating the recognition of cancer in complex tissue. The first step toward the creation of such CADx tools requires mechanisms for automatically learning to classify tissue types—a key step on the diagnosis process. Here we demonstrate that genetics-based machine learning (GBML) can be used to approach such a problem. However, to efficiently analyze this problem there is a need to develop efficient and scalable GBML implementations that are able to process very large data sets. In this paper, we propose and validate an efficient GBML technique—NAX—based on an incremental genetics-based rule learner. NAX exploits massive parallelisms via the message passing interface (MPI) and efficient rule-matching using hardware-implemented operations. Results demonstrate that NAX is capable of performing prostate tissue classification efficiently, making a compelling case for using GBML implementations as efficient and powerful tools for biomedical image processing.

X. Llorà (✉)

National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign,
1205 W. Clark Street, Urbana, IL 61801, USA
e-mail: xllora@uiuc.edu

A. Priya · R. Bhargava

Department of Bioengineering, University of Illinois at Urbana-Champaign, 1304 W. Springfield Ave.,
Urbana, IL 61801, USA

A. Priya

e-mail: priya@uiuc.edu

R. Bhargava

e-mail: rxb@uiuc.edu

R. Bhargava

Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign,
405 N. Mathews Ave., Urbana, IL 61801, USA

Keywords Observer-invariant histopathology · Genetics-based machine learning · Learning Classifier Systems · Hardware acceleration · Vector instruction · SSE2 · MPI · Massive parallelism

1 Introduction

Pathologist opinion of structures in stained tissue is the definitive diagnosis for almost all cancers and provides critical input for therapy. In particular, prostate cancer accounts for one-third of noncutaneous cancers diagnosed in US men. Hence, it is, appropriately, the subject of heightened public awareness and widespread screening. If prostate-specific antigen (PSA) or digital rectal screens are abnormal, a biopsy is needed to definitively detect or rule out cancer. Pathologic status of biopsied tissue not only forms the definitive diagnosis but constitutes an important cornerstone of therapy and prognosis. There is, however, a need to add useful information to diagnoses and to introduce new technologies that allow economical cancer detection to focus limited healthcare resources. In pathology practice, widespread screening results in a large workload of biopsied men, in turn, placing a increasing demand on services. Operator fatigue is well documented and guidelines limit the workload and rate of examination of samples by a single operator. Importantly, newly detected cancers are increasingly moderate grade tumors in which pathologist opinion variation complicates decision-making.

For the reasons above, there is an urgent need for automated and objective pathology tools. We have sought to address these requirements through novel Fourier transform infrared (FTIR) spectroscopy-based, computer-aided diagnoses for prostate cancer and develop the required microscopy and software tools to enable its application. FTIR spectroscopic imaging is a new technique that combines the spatial specificity of optical microscopy and the biochemical content of spectroscopy. As opposed to thermal infrared imaging, FTIR imaging measures the absorption properties of tissue through a spectrum consisting of (typically) 1024–2048 wavelength elements per pixel. Since IR spectra reflect the molecular composition of the tissue, image contrast arises from differences in endogenous chemical species. As opposed to visible microscopy of stained tissue that requires a human eye to detect changes, numerical computation is required to extract information from IR spectra of unstained tissue. Extracted information, based on a computer algorithm, is inherently objective and automated (Lattouf and Saad 2002; Fernandez et al. 2005; Levin and Bhargava 2005; Bhargava et al. 2006).

Uniting spectroscopic imaging data and computer-aided diagnoses (CADx), we seek to provide a new approach to pathology by automating the recognition of cancer in complex tissue. This is an exciting paradigm in which disease diagnoses are objective and reproducible; yet do not require any specialized reagents or human intervention. The first step toward the creation of such CADx tools requires mechanisms for reliable and automated tissue type classification. In this paper we demonstrate how genetics-based machine learning tools can achieve such a goal. Interpretability of the learned models and efficient processing of very large data sets have lead us to rule-based models—easy to interpret—and genetics-based machine learning—inherent massively parallel methods with the required scalability properties to address very large data sets. We present the method and the efficiency enhancement techniques proposed to address automated tissues classification. When pushed beyond the relatively small problems traditionally used to test such methods, an need for efficient and scalable implementations becomes a key research topic

that needs to be addressed. We designed the proposed a technique with such constraints in mind. A modified version of an incremental genetics-based rule learner that exploits massive parallelisms—via the message passing interface (MPI)—and efficient rule-matching using hardware-oriented operations. We name this system NAX. NAX is compared to traditional and genetics-based machine learning techniques on an array of publicly available data sets. We also report the initial results achieved using the proposed technique when classifying prostate tissue.

The remainder of the paper is structured as follows. We present an overview of the problem addressed in Sect. 2, paying special attention to tissue classification. We discuss in Sect. 3 the hurdles that traditional genetics-based machine learning implementations face when applied to very large data sets. Section 4 presents our solution to those hurdles. We also describe the incremental rule learner proposed for tissue classification. Last, we summarize results on publicly-available data sets and the preliminary results for tissue classification on a prostate tissue microarray in Sect. 5. Finally, in Sect. 6, we present conclusions and further work.

2 Biomedical imaging and data mining

This section presents an overview of the problem addressed in this paper. We first introduce infrared spectroscopic imaging as a potentially powerful tool for cancer diagnosis and prognosis. Then, we explore the protocols that provide raw high-quality data that for data mining. Finally, we conclude by focusing on the key task, tissue classification, by focusing on prostate tissue.

2.1 Infrared spectroscopy and imaging for cancer diagnosis and prognosis

Infrared spectroscopy is a well-established molecular technique and is widely used in chemical analyses. The fundamental principle governing the response of any material is that the vibrational modes of molecules are resonant in energy with photons in the mid-infrared region (2–14 μm) of the electromagnetic spectrum. Hence, when photons of energy that are resonant with the material's molecular composition are incident, a number are absorbed. The number absorbed is directly proportion to the number of chemical species that are excited. Hence, any material has a characteristic frequency-dependent absorption profile called a spectrum. An infrared spectrum is often termed the “optical fingerprint” of a material as it can help uniquely identify molecular composition—see Fig. 1.

Researchers, including us, have contributed to develop an imaging version of spectroscopy that is essentially similar to an optical microscope. In this mode of spectroscopy, images are acquired in the manner of optical microscopy with one important difference. Instead of measuring the intensity of three colors for a visible image, several thousand intensity values are acquired at each pixel in the image as a function of wavelength (spectrum at each pixel). The resulting data set is three dimensional (2 spatial and 1 spectral indices) consisting typically of a size $256 \times 256 \times 1024$, but extending to sizes such as $3500 \times 3500 \times 2048$. Since each data point is stored as a 16-bit number, the data size typically runs into several tens to hundreds of gigabytes.

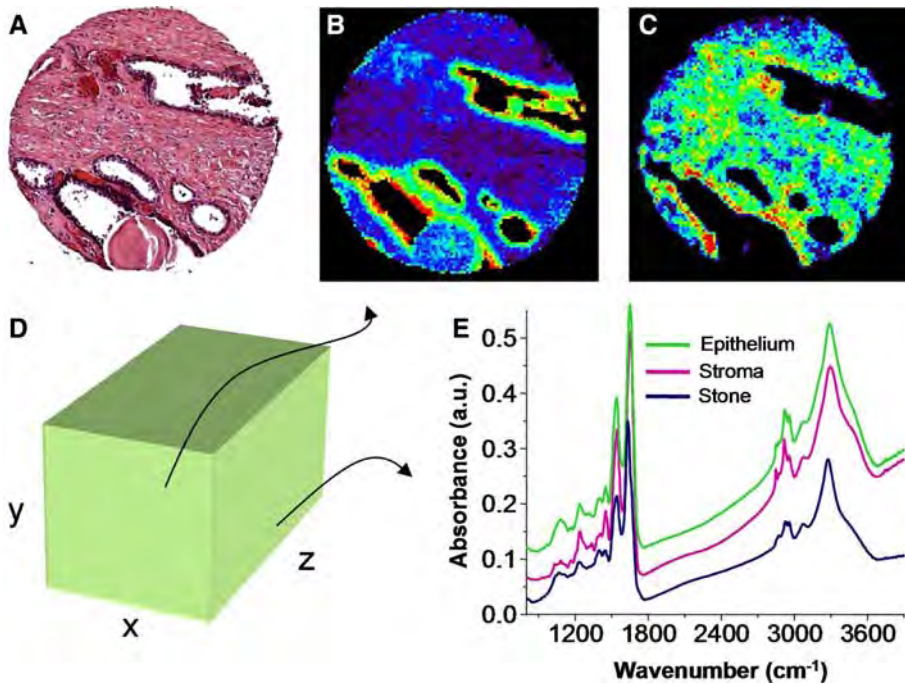


Fig. 1 Conventional staining and automated recognition by chemical imaging. (A) Typical H&E stained sample, in which structures are deduced from experience by a human. Highlights of specific regions in the manner of H&E is possible using FTIR imaging without stains. (B) Absorption at 1080 cm^{-1} commonly attributed to nucleic acids and (C) to proteins of the stroma. The data obtained is 3 dimensional (D) from which spectra (E) or images at specific spectral features may be plotted

2.2 Mining the spectra: Two sequential problems

Though the continued development of fast FTIR microspectroscopy represents an exciting opportunity for pathology, handling the resultant data and rapidly providing classifications remains a critical challenge. First, the sheer volume of data—potentially larger than 10 GB a day—represents an organizational and retrieval challenge. Next, extraction of useful information in short time periods requires the formulation of optimal protocols. Third, the automated cancer segmentation problem is very complex and offers a number of routes and levels of data that need to be analyzed to determine the optimal approach for application in a laboratory.

The typical application is the need to extract information from the data set such that it is clinically relevant. Hence, the output of the data mining algorithm to be developed is well-bounded and clearly defined. For example, in the prostate there are two levels of interest. In the first level, the pathologist examines the tissue to determine if there are any epithelial cells. Since more than 95% of prostate cancers arise in epithelial cells, transformations in this class of cells forms the diagnostic basis and a primary determinant of therapy. Other cell types of interest are lymphocytes that may indicate inflammation, blood vessel density that may indicate the development of new blood supply indicative of cancer growth and nerves that may be invaded by cancer cells. Hence, any automated approach to pathology must first identify cell types accurately. The second step in pathology follows. Once

epithelial cells are located, their spatial patterns are indicative of disease states. In our imaging approach, we can identify both spatial patterns as well as chemical patterns in epithelial cells. Hence, the task would be to use either or both to classify disease. In this paper, we focus only on the accurate identification/classification of tissue types as the first step of the path that leads to obtaining the correct pixels of epithelium.

2.3 Tissue classification for prostate arrays

Prostate tissue is structurally complex, consisting primarily of glandular ducts lined by epithelial cells and supported by heterogeneous stroma. This tissue also contains blood vessels, blood, nerves, ganglion cells, lymphocytes and stones (which are comprised of luminal secretions of cellular debris) that organize into structure measuring from tens to hundreds of microns. These structures are readily observable within stained tissue using bright-field microscopy at low to medium magnifications. Hence, in applying FTIR imaging (Levin and Bhargava 2005), we obtain the common structural detail employed clinically and, additionally, spectral information indicative of tissue biochemistry. As histologic classes contain identical chemical components, infrared vibrational spectra are similar but reveal small differences in specific absorbance features. The technique proposed by Fernandez et al. (2005) examines each cell types' spectra and transforms each spectrum into a vector of describing features—usually around the hundreds. A complete description of this process is beyond the scope of this paper and can be found elsewhere (Fernandez et al. 2005). Each pixel (cell present in the slice of micro array under analysis) has an assigned spatial position in the array while the tissue type is assigned by a highly experienced pathologist. Thus, the tissue classification can be cast into a supervised classification problem (Mitchell 1997), where all the attributes are real-valued and the class is the tissue type—ten classes: *epithelium*, *fibrous stroma*, *mixed stroma*, *muscle*, *stone*, *lymphocytes*, *endothelium*, *nerve*, *ganglion*, and *blood*. Figure 2 presents tissue types that can be assigned by examining a stained image obtained, after the FTIR microspectroscopy on unstained tissue, by the pathologist. Each marked pixel in the image becomes a training example; hence, the usual smallest data set is around hundreds of thousand records per array.

3 Larger, bigger, and faster genetics-based machine learning

Bernadó et al. (2001) presented a first empirical comparison between genetics-based machine learning techniques (GBML) and traditional machine learning approached. The authors reported that GBML techniques were as competent as traditional techniques. Later, Bacardit and Butz (2006) repeated the analysis, obtaining similar results. Most of the experiments presented on both papers used publicly available data sets provided by the *University of California at Irvine* repository (Merz and Murphy 1998). Most of the data sets are defined over tens of features and up to few thousands of records—in the larger cases. However, a key property of GBML approaches is its intrinsic massive parallelism and scalability properties. Cantú-Paz (2000) presented how efficient and accurate genetics algorithms could be assembled, and Llorà (2002) presented how such algorithms can be efficiently used for machine learning and data mining. However, there are elements that need to be revisited when we want to efficiently apply GBML techniques to large data sets such as the one described in the previous section.

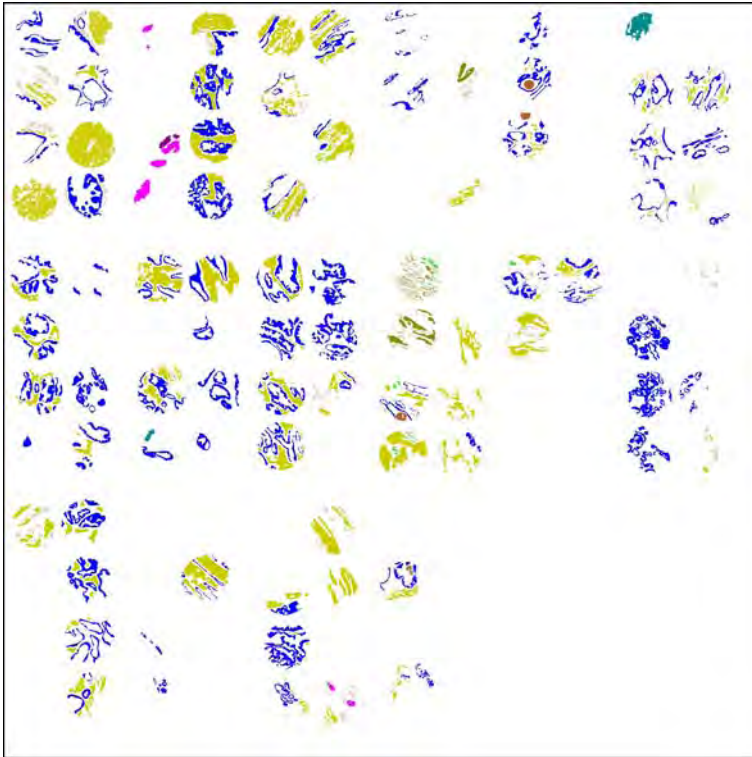


Fig. 2 The figure presents the tissue labeling provided by a pathologist biopsy section of human prostate tissue. Each spot represents the section of a needle. Different colors represent different tissue types

The GBML techniques require evaluating candidate solutions against the original data set matching the candidate solutions (e.g., rules, decision trees, prototypes) against all the instances in the data set. Regardless of the flavor used, Llorà and Sastry (2006) showed that, as the problem grows, rule matching governs the execution time. For small data sets (teens of attributes and few thousands of records) the matching process takes more than 85% of the overall execution time marginalizing the contribution of the other genetic operators. This number increases to 98% and above, when we move to data sets with few hundreds of attributes and few hundred thousands of records. More than 98% of the time is spent evaluating candidate solutions. Each evaluation can be computed in parallel. Moreover, the evaluation process may also be parallelized on very large data sets by splitting and distributing the data across the computational resources. A detailed description of the parallelization alternatives of GBML techniques can be found elsewhere (Llorà 2002).

Currently available off-the-shelf GBML methods and software distributions (Barry and Drugow-itsch 1997; Llorà 2006) do not usually target large data sets. The two main bottlenecks are large memory footprints and sequential-processing oriented processes. Generally speaking, they were designed to run on single processor machines with enough memory to fit the entire data set. Hence, designers did not paying much

attention to the memory footprint required to store the data set—usually completely loaded into memory and the population of candidate solutions. These large complex structures were geared to facilitate the programming effort, but they are not designed toward the efficient evaluation of the candidate solutions. However, efforts have been made to push GBML methods into domains which require processing large data sets. Three different works need to be mentioned here. Flockhart (1995) proposed and implemented GA-MINER, one of the earliest effort to create data mining systems based on GBML systems that scale across symmetric multi-processors and massively parallel multi-processors. Flockhart (1995) reviewed different encoding and parallelization schemes and conducted proper scalability studies. Llorà (2002) explored how fine-grained parallel genetic algorithms could become efficient models for data mining. Theoretical analysis of performance and scalability were developed and validated with proper simulations. Recently, Llorà and Sastry (2006) explored how current hardware can efficiently speed up rule matching against large data sets. These three approaches are the basis of the incremental rule learning proposed in the next section to approach very large data sets.

Another important issue in real-world problems is the class distribution. Usually most real problems have a clear class imbalance. Recently, Orriols-Puig and Bernadó-Mansilla (2006) have revisited this issue, showing how GBML techniques successfully learn and maintain proper descriptions for those minority classes. If not designed properly, descriptions of majority classes will tend to govern the learned models, starving the description of minority classes. Prostate tissue classification is a clear example of extreme class imbalance. Figure 3 presents the tissue type class distribution. The smaller tissue type has 64 records, where as the larger classes have several tens of thousands records. hence, the developed approaches must account for class size variation.

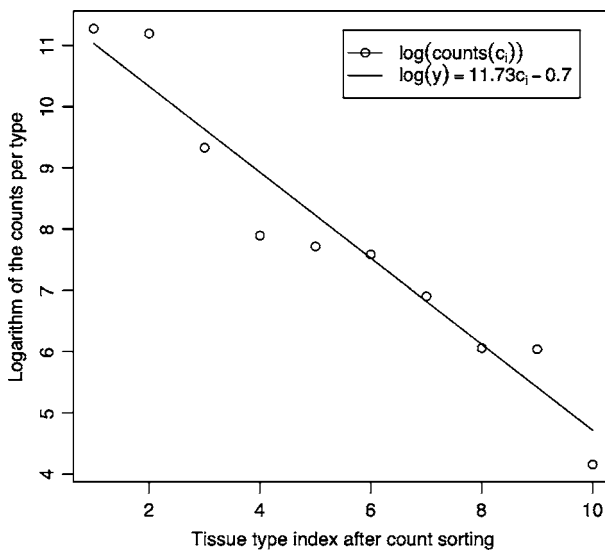


Fig. 3 Figure shows the tissue class distribution. Once the classes are reordered according to their frequency in the data set, we can easily appreciate the extreme imbalance—the smaller tissue type has 64 records, where as the larger classes have several tens of thousands records

4 The road to tractability

We describe in this section the steps we took to design a GBML method (NAX) able to deal with very large data sets with class imbalance. NAX evolves, one at a time, maximally general and maximally accurate rules. Then, the covered instance are removed and another maximally general and maximally general rule is evolved and added to the previously stored one forming a decision list. This process continues until no uncovered instances are left—this process is also referred as the sequential covering procedure (Cordón et al. 2001). Llorà et al. (2005) showed that maximally general and maximally accurate rules (Wilson 1995) could also be evolved using Pittsburgh-style Learning Classifier Systems. Later, Llorà et al. (2007) showed that competent genetic algorithms (Goldberg 2002) evolve such rules quickly, reliably, and accurately. The rest of this section describes (1) efficient implementation techniques to deal with very large data sets, (2) the impact of class imbalance, and (3) the NAX algorithm proposed.

4.1 Efficient implementations

As introduced earlier, when dealing with very large data sets, and regardless of the flavor of the GBML technique used, we may spend up to 98% of the computational cycles trying to match rules to the original data set (Llorà and Sastry 2006). Each solution evaluation is independent of each other and, hence, it can be computed in parallel. Moreover, even the matching nature of a rule—the representation we will use from now on—is highly parallel, since conditions require performing simultaneous checks against different attributes per record. Thus, efficient implementation can take advantage of parallelizing both elements.

4.1.1 Exploiting the hardware acceleration

Recently, multimedia and scientific applications have pushed CPU manufactures to include support for vector instructions again in their processors. Both applications areas require heavy calculations based on vector arithmetic. Simple vector operations such as *add* or *product* are repeated over and over. During 1980s and 1990s supercomputers, such as Cray machines, were able to issue hardware instructions that enabled basic vector arithmetics. A more constrained scheme, however, has made its way into general-purpose processors thanks to the push of multimedia and scientific applications. Main chip manufactures—IBM, Intel, and AMD—have introduced vector instruction sets—AltiVec, SSE3, and 3DNow⁺—that allow vector operations over packs of 128 bits by hardware. We will focus on a subset of instructions that are able to deal with floating point vectors. This subset of instructions manipulate groups of four floating-point numbers. These instructions are the basis of the fast rule matching mechanism proposed.

Our goal is to evolve a set of rules that correctly classifies the current data set from prostate tissue. Using a knowledge representation based on rules allows us to inspect the learned model, gaining insight into the biological problem as well. All the attributes of the domain are real-value and the conditions of the rules need to be able to express conditions in a \mathbb{R}^n spaces. We use a similar rule encoding to the one proposed by Wilson (2000b)—a variation of the original work proposed by Wilson (2000a) and later reviewed by Stone and Bull (2003)—and widely used in the GBML community. Rules express the conjunction of tests across attributes. Each test may be defined in multiple flavors but, without loss of

generality, we picked a simple interval based one. A simple example of an *if-then* rule, could be expressed as follows:

$$1.0 \leq a_0 \leq 2.3 \wedge \dots \wedge 10.0 \leq a_n \leq 23 \rightarrow c_1 \quad (1)$$

Where the condition is the conjunction of the different attribute tests and the outcome is the predicted class—a tissue type. We also allow a special condition—don't care—which just always returns true, allowing condition generalization. The rule below illustrates an example of a generalized rule.

$$1.0 \leq a_0 \leq 2.3 \wedge -3.0 \leq a_3 \leq 2 \rightarrow c_1 \quad (2)$$

All attributes except a_0 and a_3 were marked as don't care.

Each condition can be encoded using 2 floating-point numbers per condition, where α_i contains the lower bound of the condition and ω_i its upper bound. Thus, the condition $\alpha_i \leq a_0 \leq \omega_i$ just requires to store the two floating-point numbers. For efficiency reasons we store them in two separate vectors, one containing the lower bounds and the other containing the upper bounds. The position in a vector indicates the attribute being tested. The don't care condition is simply encoded as $\alpha_i > \omega_i$ and, hence, we do not need to store any extra information.

Matching a rule requires performing the individual condition tests before the final *and* operation can be computed. Vector instruction sets improve the performance of this process by performing four operations at once. Actually, this process may be regarded as four parallel running pipelines. The process can be further improved by stopping the matching process when one test fails—since that will turn the condition into false.

Figure 4 presents a C implementation the proposed hardware-supported rule matching. The code assumes that the two vectors containing the upper and lower bounds are provided and records are stored in a two dimensional matrix. Figure 5 presents the vectorized implementation of the code presented in Fig. 4 using SSE2 instructions. Exploiting the hardware available can speed between 3 and 3.5 times the matching process, as also shown elsewhere (Llorà and Sastry 2006).

4.1.2 Massive parallelism

Since most of the time is spent on the evaluation of candidate rules when dealing with large data sets, our next goal was to find a parallelization model that could take advantage of this peculiarity. Due the quasi embarrassing parallel (Grama et al. 2003) nature of the candidate rule evaluation, we designed a coarse-grain parallel model for distributing the evaluation load. Cantú-Paz (2000) proposed several schemes, showing the importance of the trade-off between computation time and time spent communicating. When designing the parallel model, we focused on minimizing the communication cost. Usually, a feasible solution could be a master/slave one—the computation time is much larger than the communication time. However, GBML approaches tend to use rather large populations, forcing us to send rule sets to the evaluation slaves and collect the resulting fitness. These schemes also increment the sequential sections that cannot be parallelized, threatening the overall speedup of the parallel implementation as a result of Amdahl's law (Amdahl 1967).

To minimize such communication cost, each processor runs an identical NAX algorithm. They are all seeded in the same manner, hence, performing the same genetic operations and only differing in the portion of the population being evaluated. Thus, the population is

```

1. void match_seq_rule_set ( RuleSet * rs, InstanceSet is, int iDim, int iRows ) {
2.     int i,j,k,iCnt,iClsIdx,iGround,iPred;
3.     register int iMatcheable;
4.     Instance ins;
5.
6.     iClsIdx = rs->iCorrectedDim;
7.     clean_fitness_rules_set(rs);
8.     for ( i=0 ; i<iRows ; i++ ) {
9.         ins = is[i];
10.        iPred=-1;
11.        for ( j=0 ; iPred==-1 && j<rs->iLen ; j++ ) {
12.            iMatcheable = 1;
13.            for ( iCnt=0,k=j*(rs->iCorrectedDim+VBSIF) ;
14.                iMatcheable && k<j*(rs->iCorrectedDim+VBSIF)+rs->iDim ;
15.                k++,iCnt++ ) {
16.                iMatcheable = iMatcheable &&
17.                    !( (rs->pfLB[k]<=rs->pfUB[k]) &&
18.                      ( ins[iCnt]<rs->pfLB[k] || ins[iCnt]>rs->pfUB[k])));
19.            }
20.            if ( iMatcheable )
21.                iPred = rs->pfLB[j*(rs->iCorrectedDim+VBSIF)+rs->iCorrectedDim];
22.        }
23.        iPred = (iPred==-1)?rs->iClasses:iPred;
24.        iGround=(int)ins[iClsIdx];
25.        rs->pConfMat[iGround][iPred]++;
26.    }
27. }

```

Fig. 4 This figure presents a sequential implementation of the rule matched process in C. A rule set is match against a data set. Lines 16, 17, and 18 implement the condition test for one attribute. The implementation also computes the confusion matrix that contains the ground truth versus predicted class

treated as collection of chunks where each processor evaluates its own assigned chunk, sharing the fitness of the individuals in its chunk with the rest of the processors. Fitness can be encapsulated and broadcasted maximizing the occupation of the underlying packing frames used by the network infrastructure. Moreover, this approach also removes the need for sending the actual rules back and forth between processors—as a master/slave approach would require—thus, minimizing the communication to the bare minimum—the fitness. Figure 6 presents a conceptual scheme of the parallel architecture of NAX.

To implement the model presented in Fig. 6, we used C and a *message passing interface* (MPI)—we used the OpenMPI implementation (Gabriel et al. 2004). Figure 7 shows the code in charge of the parallel evaluation. Each processor computes which individuals are assigned to it. Then it computes the fitness and, finally, it just broadcast the computed fitness. The rest of the process is left untouched, and besides the cooperative evaluation, all the processors end generating the same evolutionary trace.

4.2 Rule sets as individuals

One main characteristic of the so-called Pittsburgh-style learning classifier systems—a particular type of GBML—is that individuals encode a rule set (Goldberg 1989; Llorà and Garrell 2001; Goldberg 2002). Thus, evolutionary mechanisms directly recombine one rule set against another one. For classification tasks of moderate complexity, the rule sets are

```

1. #define VEC_MATCH(vecFLB,fLB,vecFUB,fUB,vecINS,fIN,vecTmp,vecOne,vecRes) {\
2.     vecFLB = _mm_load_ps(fLB);\
3.     vecFUB = _mm_load_ps(fUB);\
4.     vecINS = _mm_load_ps(fIN);\
5.     \
6.     vecRes = (__m128i)_mm_cmpgt_ps(vecFUB,vecFLB);\
7.     vecTmp = _mm_or_si128(\
8.         (__m128i)_mm_cmpgt_ps(vecFLB,vecINS),\
9.         (__m128i)_mm_cmpgt_ps(vecINS,vecFUB)\
10.    );\
11.     vecRes = _mm_andnot_si128(_mm_and_si128(vecRes,vecTmp),vecOne);\
12. }
13.
14. void match_rule_set ( RuleSet * rs, InstanceSet is, int iDim, int iRows ) {
15.     int i,j,k,iCnt,iClsIdx,iGround,iPred;
16.     register int iMatcheable;
17.     Instance ins;
18.
19.     __m128i vecRes,vecTmp,vecOne;
20.     __m128 vecFLB,vecFUB,vecINS;
21.
22.     vecOne = (__m128i){-1,-1};
23.
24.     iClsIdx = rs->iCorrectedDim;
25.     clean_fitness_rules_set(rs);
26.     for ( i=0 ; i<iRows ; i++ ) {
27.         // Classify the instance
28.         ins = is[i];
29.         iPred=-1;
30.         for ( j=0 ; iPred==-1 && j<rs->iLen ; j++ ) {
31.             iMatcheable = 1;
32.             for ( iCnt=0,k=j*(rs->iCorrectedDim+VBSIF) ;
33.                 iMatcheable && k<j*(rs->iCorrectedDim+VBSIF)+rs->iDim ;
34.                 k+=VBSIF,iCnt+=VBSIF ) {
35.                 VEC_MATCH(vecFLB,&(rs->pfLB[k]),
36.                     vecFUB,&(rs->pfUB[k]),
37.                     vecINS,&(ins[iCnt]),vecTmp,vecOne,vecRes);
38.                 iMatcheable = 0xFFFF==_mm_movemask_epi8(vecRes);
39.             }
40.             if ( iMatcheable )
41.                 iPred = rs->pfLB[j*(rs->iCorrectedDim+VBSIF)+rs->iCorrectedDim];
42.             iPred = (iPred==-1)?rs->iClasses:iPred;
43.             iGround=(int)ins[iClsIdx];
44.             rs->pConfMat[iGround][iPred]++;
45.         }
46.     }

```

Fig. 5 This figure presents a vectorized implementation of the rule matching process presented in Fig. 4. Lines 1–12 implement the parallelized test against four attributes using vector instructions. The code is written using C intrinsics for SSE2 compatible architectures. This code runs on P4 or newer Intel processors and Opteron or Athlon 64 AMD processors

not large. However, for complex problems, the potential number of required rules to ensure proper classification may need large amounts of memory that become prohibitive. The requirements increase even further in the presence of noise (Llorà and Goldberg 2003).

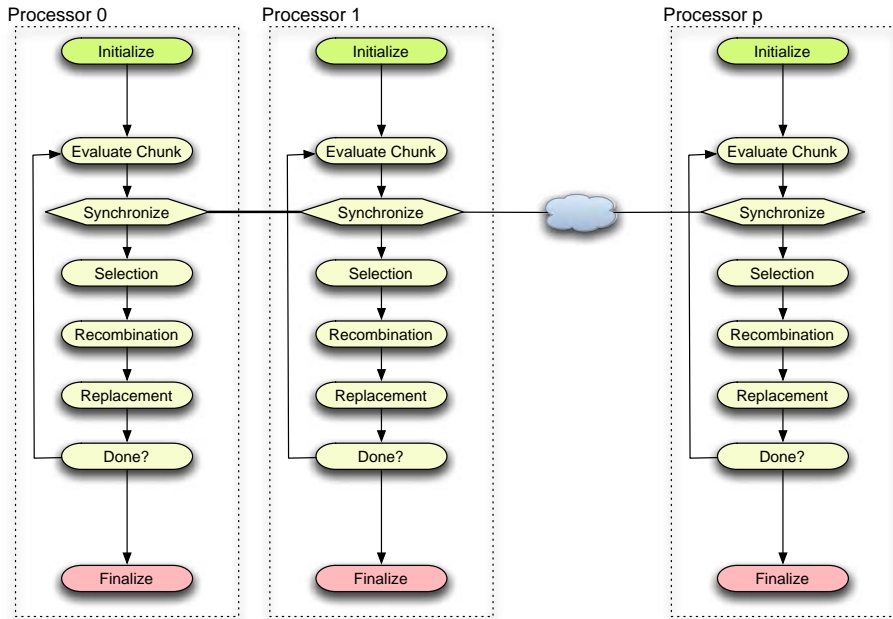


Fig. 6 This figure illustrates the parallel model implemented. Each processor is running the same identical NAX algorithm. They only differ in the portion of the population being evaluated. The population is treated as collection of chunks where each processor evaluates its own assigned chunks sharing the fitness of these individuals with the rest of the processors. This approach minimizes the communication cost

Parallelization may not help much if we need to send large rule sets across the communication network. For such reasons, GBML techniques work very well on moderate complexity problems (Bernadó et al. 2001; Bacardit and Butz 2006). However, they need to be modified to deal with complex and large data set, and also avoid the boundaries imposed by the issues mentioned above.

4.3 NAX: Incremental rule learning for very large data sets

An incremental rule learning approach may alleviate memory footprint requirements by evolving only one rule at a time, hence, reducing the memory requirements. However, one rule by itself cannot solve complex problems. For such a reason, each evolved rule is added to the final rule set, and the covered examples are removed from the current training set. The process is repeated until no instances are left in the training set. This approach already introduced by Cordon et al. (2001) and later also used by Bacardit and Krasnogor (2006) allows maintaining relatively small memory footprints, making feasible processing large data sets—as the prostate tissue classification data set. However, an incremental approach to the construction of the rule set requires paying special attention to the way rules are evolved. For each run of the genetic algorithm used to evolve a rule, we would like to obtain a maximally general and maximally accurate rule, that is, a rule that covers the maximum number of example without making mistakes (Wilson 1995).

```

1. void evaluate_population ( Population * pp, InstanceSet is, int iDim, int iRows )
2. {
3.     int i;
4.
5.     /* Compute the fragments of this processor */
6.     int iFrag = pp->iLen/FCS_processes;
7.     int iInit = FCS_process_id*iFrag;
8.     int iLast = (FCS_process_id+1==FCS_processes)?
9.                 pp->iLen:
10.                (FCS_process_id+1)*iFrag;
11.     int iCnt = 0;
12.     int j,k,l;
13.
14.     /* Create the bucket for the broadcast */
15.     float faFit[2*iFrag];
16.     float faTmp[2*iFrag];
17.
18.     /* Evaluate the given chunk assigned to the processor */
19.     for ( i=iInit,iCnt=0 ; i<iLast ; i++,iCnt++ ) {
20.         match_rule_set(pp->prs[i],is,iDim,iRows );
21.         compute_raw_accuracy_fitness_rule_set(pp->prs[i]);
22.         faFit[iCnt] = pp->prs[i]->fFitness;
23.     }
24.
25.     /* Broadcast each of the chunks */
26.     for ( i=0 ; i<FCS_processes ; i++ ) {
27.         MPI_Bcast((i==FCS_process_id)?faFit:faTmp,iCnt,MPI_FLOAT,i,MPI_COMM_WORLD);
28.         if ( i!=FCS_process_id )
29.             for ( l=0,j=i*iFrag, k=(i+1)*iFrag ; j<k ; j++,l++ )
30.                 pp->prs[j]->fFitness = faTmp[l];
31.     }
32. }

```

Fig. 7 This figure presents an implementation of the proposed parallel evaluation scheme using C and MPI. The piece of code presented below is the only one modified to provide such parallelization capabilities. Each processor computes which individuals are assigned to it (lines 6–10), then it computes the fitness (lines 10–23), and then it just broadcast the computed fitness (lines 26–31)

Llorà et al. (2007) have shown that evolving such rules is possible. In order to promote maximally general and maximally accurate rules à la XCS (Wilson 1995), we compute the *accuracy* (α) and the *error* (ε) of a rule (Llorà et al. 2005). The *accuracy* is the proportion of overall examples correctly classified, and the *error* is the proportion of incorrect classifications issued. For simplicity reasons, we use the proportion of correctly issues classifications instead, simplifying the final fitness calculation. Let n_{t+} be the number of positive examples correctly classified, n_{t-} the number of negative examples correctly classified, n_m the number of times a rule has been matched, and n_t the number of examples available. Using these values, the *accuracy* and *error* of a rule r can be computed as:

$$\alpha(r) = \frac{n_{t+}(r) + n_{t-}(r)}{n_t} \quad (3)$$

$$\varepsilon(r) = \frac{n_{t+}(r)}{n_m(r)} \quad (4)$$

Once the *accuracy* and *error* of a rule are known, the fitness can be computed as follows.

$$f(r) = \alpha(r) \cdot \varepsilon(r)^\gamma \quad (5)$$

where γ is the error penalization coefficient. The above fitness measure favors rules with a good classification accuracy and a low error, or maximally general and maximally accurate rules. By increasing γ , we can bias the search towards correct rules. This is an important element because assembling a rule set based on accurate rules guarantees the overall performance of the assembled rule set. In our experiments, we have set γ to 18 to strongly bias the search toward maximally general and maximally accurate rules.

NAX's efficient implementation of the evolutionary process is based on the techniques described using hardware acceleration—Sect. 4.1.1—and coarse-grain parallelism—Sect. 4.1.2. The genetic algorithm used was a modified version of the *simple genetic algorithm* (Goldberg 1989) using tournament selection ($s = 4$), one point crossover, and mutation based on generating new random boundary elements.

5 Experiments

This section presents the results achieved using NAX. To allow the reader to compare with other techniques, we compare the results obtained using NAX on small data sets provided by the UCI repository (Merz and Murphy 1998) to other well-known supervised learning algorithms. Finally, we present the first results on the prostate tissue prediction obtained using NAX. Results focus on the viability of the NAX approach.

5.1 Some UCI repository data sets

The UCI repository (Merz and Murphy 1998) provides several data sets for different machine learning problems. These data sets have been widely used to test traditional machine learning and GBML techniques. Table 1 lists the data sets used. Due to the nature of the prostate tissue type classification, we only chose data sets with numeric attributes. Three of these data sets are of relevant interest: (1) *son*, by far the one with larger dimensionality, (2) *gls*, the one with large number of classes, (3) *tao*, proposed by Llorà and Garrell (2001), having complex and non-linear boundaries.

Table 1 Summary of the data sets used in the experiments

ID	Data set	Size	Missing values(%)	Numeric attributes	Nominal attributes	Classes
bre	<i>Wisconsin Breast Cancer</i>	699	0.3	9	–	2
bpa	<i>Bupa Liver Disorders</i>	345	0.0	6	–	2
gls	<i>Glass</i>	214	0.0	9	–	6
h – s	<i>Heart Stats-Log</i>	270	0.0	13	–	2
ion	<i>Ionosphere</i>	351	0.0	34	–	2
irs	<i>Iris</i>	150	0.0	4	–	3
son	<i>Sonar</i>	208	0.0	60	–	2
tao	<i>Tao</i>	1888	0.0	2	–	2
win	<i>Wine</i>	178	0.0	13	–	3

Table 2 Experimental results: percentage of correct classifications and standard deviation from stratified ten-fold cross-validation runs

ID	0-R	C4.5	NAX
bre	65.52 \pm 1.16	95.42 \pm 1.69	96.43 \pm 1.72
bpa	57.97 \pm 1.23	65.70 \pm 3.84	64.07 \pm 8.36
gls	35.51 \pm 4.49	65.89 \pm 10.47	68.02 \pm 8.69
h - s	55.55 \pm 0.00	76.30 \pm 5.85	75.56 \pm 9.39
ion	64.10 \pm 1.19	89.74 \pm 5.23	89.19 \pm 5.27
irs	33.33 \pm 0.00	95.33 \pm 3.26	94.67 \pm 4.98
son	53.37 \pm 3.78	71.15 \pm 8.54	73.62 \pm 9.72
tao	49.79 \pm 0.17	95.07 \pm 2.11	97.41 \pm 0.92
win	39.89 \pm 3.22	93.82 \pm 2.85	94.34 \pm 6.09

Paired *t*-test comparisons showed no statistically significant differences between C4.5 and NAX results

0-R result are just provided as guiding base line

We could have chosen complex algorithms as baselines for NAX. However, we would not be able to use them to repeat the experimentation on the prostate tissue classification domain. The algorithms used in the comparison presented in Table 2 were 0-R (Holte 1993) (a simple base line based on majority class classification) and C4.5 (Quinlan 1993). Results show percentage of correct classifications and standard deviation from stratified ten-fold cross-validation runs. Paired *t*-test comparisons showed no statistically significant differences between the pruned tree produced by C4.5 and NAX results. This experiments also helped validate the distributed implementation proposed by NAX. Further results on empirical comparisons can be found elsewhere (Bernadó et al. 2001; Bacardit and Butz 2006).

5.2 Prostate tissue classification

With the previous results at hand, we ran NAX against the prostate tissue classification data set. The original data set is defined by 93 attributes. In this paper, however, we used the reduced version of this data set proposed by (Fernandez et al. 2005) which contains 20 selected attributes out of the 93 available. The dataset is form by 171,314 records. Our goal was to explore how well NAX could generalize over unseen tissue—this is the first step to be able to address the cancer prediction problem. The other reason that motivated such experimentation was to achieve similar accuracy results as the ones published earlier by Fernandez et al. (2005) using a modified Bayes technique. If NAX could perform at the same level, we will also obtain a set of rules of interest to the spectroscopist. The interpretation of the rules will provide insight on how to interpret the models provided by NAX—which could not be done with the models early used by Fernandez et al. (2005).

We conducted stratified 10-fold cross-validation experiments to measure the generalization capabilities of NAX for this problem. Since the problem was rather small—larger data set are being prepared to be run at the supercomputing facilities provided by the National Center for Supercomputing Applications—we run the ten-fold cross-validation runs in a 3GHz dual core Pentium D computer with 4 GB of RAM. NAX took advantage of the hardware support to speedup the matching process and uses two MPI processes to parallelize—as introduced in Fig. 6—the evaluation of the overall population. Each fold

took about one hour to complete, with the entire classification lasting less than half a day. We conducted a simple test of adding a second computer with an identical configuration. The overall time for cross-validation was reduced to half. Rough estimates—which will better measured when larger experiments are conducted on NCSA super computers—show that the sequential portion is around 1:1000 for this small data set. Numbers get better as data set increases, which demonstrates that we will be able to process very large data sets and efficiently exploit larger numbers of processors.

We proposed another measure of effectiveness, namely how many records can be processed per second. Using a single processor with the hardware acceleration mechanisms built into NAX, and the evolved rule set formed by 1,028 rules, the average throughput was around 60,000 records per second. For the prostate tissue classification, it took less than three seconds to classify the entire data set. Once the rule set is learnt, the classification problem falls again into the category of embarrassingly parallel problems (Grama et al. 2003). Since no communication is needed, the speedup grows linearly with the number of processors added—with the proper rule set replication and data set chunking. Thus, with the dual core box used we were able to just double the throughput (120,000 records per second) by chunking the data set and use both processors.

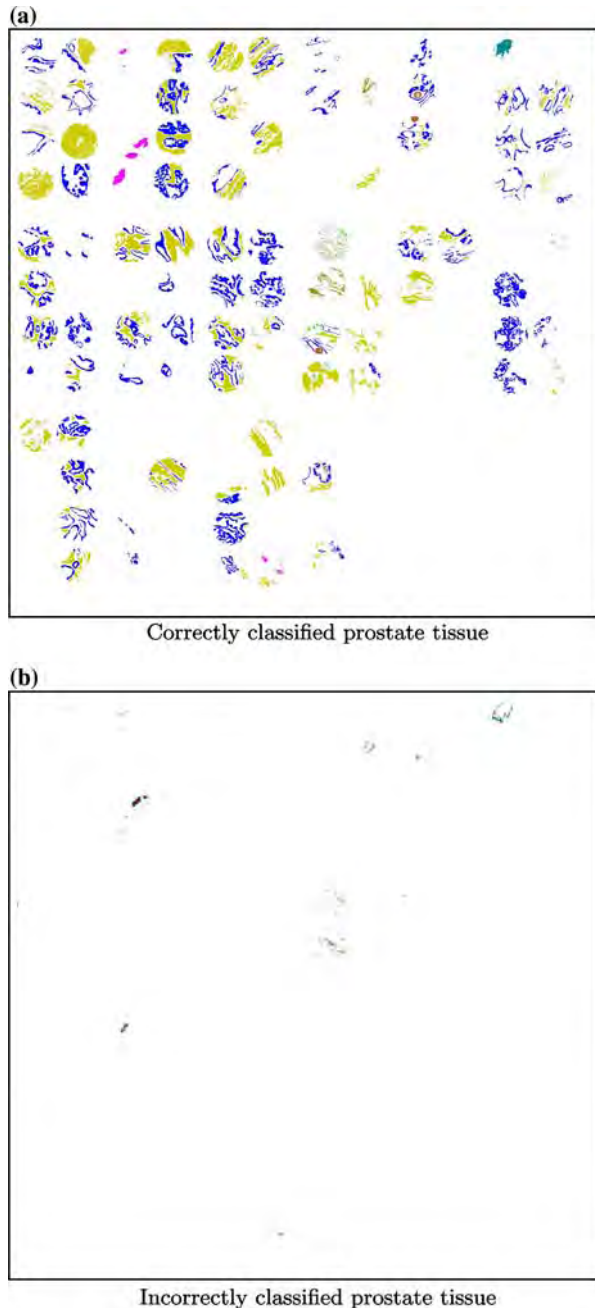
The previous results show the benefits of hardware acceleration and parallelization, but NAX was also able to achieve very competitive classification accuracy in generalization, correctly classifying 97.09 ± 0.09 of the records (pixels) during the stratified ten-fold cross-validation. Figure 8 presents the regenerated prostate tissue classification image presented in Fig. 2 using a rule set assembled by NAX. Figure 8a presents the incorrectly classified pixels. Most of the mistakes by the rule set involve similar tissues with few training records available. This trend was also shown elsewhere (Fernandez et al. 2005). C4.5 does not provide any statistically significant improvement (only a marginal, not statistically significant, 0.7%) and provided large decision trees with more than 5,000 leaves—not to mention the lack of scalability when compared to NAX.

The rule set assembled by NAX represents an incremental assembling of maximally general and maximally accurate rules. Thus, we can compute how the accuracy of such ensemble improves as new rules are added. Figure 9 presents the overall accuracy as rules are added. It shows an interesting behavior for classifying prostate tissue. Using only 20 rules out of the 1,028 evolved ones, the overall accuracy is 90%, the incorrectly classified 1.3% pixels, and 8.7% were left unclassified. After inspecting the misclassified pixels most of them belongs to borders between tissues and mislabeling arises from the image discretization—one pixel containing different tissue types. Table 3 presents the initial four rules that covering 80% of the instances belonging to the two larger tissue types—epithelium and fibrous stroma. Such results are relevant, not only for their accuracy, but also because of the insight they provide to the spectroscopist about the problem structure.

6 Conclusions and further work

This paper has presented the initial results achieved in predicting prostate tissue type using GBML techniques. Being able to classify unseen tissue quickly, reliably, and accurately, is the first step towards the creation of CADx systems that may assist a pathologist diagnosing prostate cancer. We have proposed two main efficiency enhancement techniques for GBML—exploiting hardware parallelization via vector instructions and coarse-grain parallelism via the usage of MPI libraries—which allowed us to approach very large data sets. These techniques, together with an incremental genetics-based rule learning approach to

Fig. 8 The figures presented above show the regenerated prostate tissue classification image presented in Fig. 2. **(a)** presents the correctly classified pixels. **(b)** presents the incorrectly classified pixels



assemble rule sets formed by maximally general and maximally accurate rules, have led to the creation of NAX, a system specialized on dealing with large data sets.

Results have shown accurate classification models for prostate tissue along with good scalability of the NAX implementation. Results also reveal peculiarities of the underlying problem structure. With very few rules—20—we were able to correctly classify up to 90%

Fig. 9 The rule set as a decision list. The figure presents the classification accuracy as we keep adding rules to the decision list. The first 20 initial rules are able to cover 91% of the records with a classification accuracy of 98.5–90% overall accuracy presented in the figure

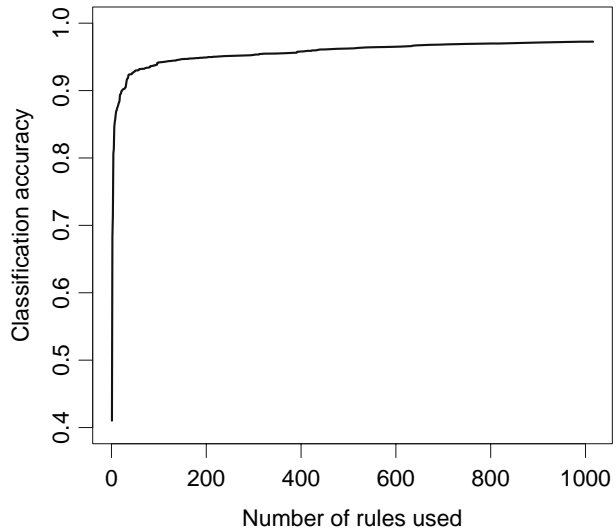


Table 3 First top four maximally general and maximally accurate rules that compose the final rule set. The rule set is treated as a decision list, thus we can easily incrementally evaluate the value of the initial four ones

Rule	Rule condition	Tissue type	Accumulated accuracy (%)	Covered records (%)
1.	$0.10 \leq a_1 \leq 0.25 \wedge 0.00 \leq a_4 \leq 0.04 \wedge$ $1.07 \leq a_8 \leq 2.01 \wedge -0.07 \leq a_{16} \leq 0.16 \wedge$ $0.25 \leq a_{17} \leq 2.86 \wedge 0.11 \leq a_{18} \leq 0.21$	→ <i>Fibrous stroma</i>	41.32	41.96
2.	$0.03 \leq a_1 \leq 0.11 \wedge 0.05 \leq a_7 \leq 0.20 \wedge$ $1231.88 \leq a_{12} \leq 1247.90 \wedge 1.98 \leq a_{17} \leq 3.83 \wedge$ $0.13 \leq a_{18} \leq 0.20$	→ <i>Epithelium</i>	68.53	69.61
3.	$0.07 \leq a_0 \leq 0.16 \wedge 0.14 \leq a_1 \leq 0.41 \wedge$ $0.71 \leq a_{10} \leq 1.13 \wedge 1527.54 \leq a_{15} \leq 1533.80 \wedge$ $0.65 \leq a_{19} \leq 1.50$	→ <i>Fibrous stroma</i>	71.59	72.75
4.	$0.05 \leq a_2 \leq 0.09 \wedge 0.76 \leq a_4 \leq 1.29 \wedge$ $1.80 \leq a_6 \leq 2.08 \wedge 0.17 \leq a_7 \leq 0.24 \wedge$ $0.26 \leq a_{16} \leq 0.53 \wedge 2.79 \leq a_{17} \leq 7.01 \wedge$ $0.21 \leq a_{18} \leq 0.32$	→ <i>Epithelium</i>	80.78	82.08

of the tissue. Our current work is focused on analyzing the larger data sets containing all the available features and different tissue sources to test the parallelization scalability of NAX on NCSA supercomputers. Once accomplished, the procedure will provide confidence in creating a CADx system to generate a diagnosis based on the evolved models.

Acknowledgments We would like to thank David E. Goldberg for his continual support and encouragement, allowing us to have access to the IlliGAL resources. Thanks also to Kumara Sastry for hallway discussions and to the Automated Learning Group and the Data-Intensive Technologies and Applications at the National Center for Supercomputing Applications for hosting this joint collaboration. This work was sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant

FA9550-06-1-0370, the National Science Foundation under grant IIS-02-09199, and the National Institute of Health. The US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research, the National Science Foundation, or the US Government. Rohit Bhargava would like to acknowledge collaborators over the years, especially Dr. Stephen M. Hewitt and Dr. Ira W. Levin of the National Institutes of Health, for numerous useful discussions and guidance. Funding for this work was provided in part by University of Illinois Research Board and by the Department of Defense Prostate Cancer Research Program. This work was also funded in part by the National Center for Supercomputing Applications and the University of Illinois, under the auspices of the NCSA/UIUC faculty fellows program.

References

- Amdahl G (1967) Validity of the single processor approach to achieving large-scale computing capabilities. In Proceedings of the American federation of information processing societies conference (AFIPS). 30:483–485 AFIPS
- Bacardit J, Butz M (2006) Advances at the frontier of Learning Classifier Systems. Chapter data mining in Learning Classifier Systems: Comparing XCS with GAssist, vol I. Springer
- Bacardit J, Krasnogor N (2006) Biohel: Bioinformatics-oriented hierarchical evolutionary learning (Nottingham ePrints). University of Nottingham
- Barry A, Drugowitsch J (1997) LCSWeb: the LCS wiki. <http://www.lcsweb.cs.bath.ac.uk/>
- Bernadó E, Llorà X, Garrell J (2001) Advances in Learning Classifier Systems: 4th international workshop (IWLCS 2001). Chapter XCS and GALE: a comparative study of two Learning Classifier Systems with six other learning algorithms on classification tasks. Springer Berlin, Heidelberg, pp 115–132
- Bhargava R, Fernandez D, Hewitt S, Levin I (2006) High throughput assessment of cells and tissues: Bayesian classification of spectral metrics from infrared vibrational spectroscopic imaging data. *Biochemica et Biophysica Acta* 1758(7):830–845
- Cantú-Paz E (2000) Efficient and accurate parallel genetic algorithms. Kluwer Academic Publishers
- Cordón O, Herrera F, Hoffmann F, Magdalena L (2001) Genetic fuzzy systems. Evolutionary tuning and learning of fuzzy knowledge bases. World Scientific
- Fernandez D, Bhargava R, Hewitt S, Levin I (2005) Infrared spectroscopic imaging for histopathologic recognition. *Nat Biotechnol* 23(4):469–474
- Flockhart I (1995) GA-MINER: parallel data mining with hierarchical genetic algorithms (final report). (Technical Report Technical Report EPCCAUKMS-GA-MINER-REPORT 1.0). University of Edinburgh
- Gabriel E, Fagg G, Bosilca G, Angskun T, Dongarra J, Squyres J, Sahay V, Kambadur P, Barrett B, Lumsdaine A, Castain R, Daniel D, Graham R, Woodall T (2004) Open MPI: goals, concept, and design of a next generation MPI implementation. In Proceedings of the 11th European PVMMPI Users' group meeting Springer
- Goldberg D (1989) Genetic algorithms in search, optimization, and machine learning. Addison-Wesley Professional
- Goldberg D (2002) The design of innovation: lessons from and for competent genetic algorithms. Springer
- Grama A, Gupta A, Karypis G, Kumar V (2003) Introduction to parallel computing. Addison-Wesley
- Holte R (1993) Very simple classification rules perform well on most commonly used datasets. *Mach Learn* 11:63–91
- Lattouf J-B, Saad F (2002) Gleason score on biopsy: is it reliable for predicting the final grade on pathology? *BJU Int* 90:694–699
- Levin I, Bhargava R (2005) Fourier transform infrared vibrational spectroscopic imaging: integrating microscopy and molecular recognition. *Annu Rev Phys Chem* 56: 429–474
- Llorà X (2002) Genetics-based machine learning using fine-grained parallelism for data mining. Doctoral dissertation, Enginyeria i Arquitectura La Salle. Ramon Llull University, Barcelona, Catalonia, European Union
- Llorà X (2006) Learning Classifier Systems and other genetics-based machine learning Blog. <http://www-illigal.ge.uiuc.edu/lcs-n-gbml/>
- Llorà X, Garrell J (2001) Knowledge-independent data mining with fine-grained parallel evolutionary algorithms. In Proceedings of the genetic and evolutionary computation conference (GECCO'2001). Morgan Kaufmann Publishers, pp 461–468

- Llorà X, Goldberg D (2003) Bounding the effect of noise in multiobjective Learning Classifier Systems. *Evol Comput J* 11(3):279–298
- Llorà X, Sastry K (2006) Fast rule matching for Learning Classifier Systems via vector instructions. In *Proceedings of the 2006 genetic and evolutionary computation conference*. ACM Press, pp 1513–1520
- Llorà X, Sastry K, Goldberg D (2005) The compact classifier system: motivation, analysis and first results. In *Proceedings of the congress on evolutionary computation*, vol 1. IEEE press, (Also as IlliGAL TR No 2005019, pp 596–603)
- Llorà X, Sastry K, Goldberg D, de la Ossa L (2007) The χ -ary extended compact classifier system: linkage learning in Pittsburgh LCS. In *Advances at the frontier of Learning Classifier Systems*, vol II. IlliGAL report no 2006015. Springer, pp (in preparation)
- Merz CJ, Murphy PM (1998) UCI repository for machine learning data-bases. <http://www.ics.uci.edu/~mllearn/MLRepository.html>
- Mitchell T (1997) *Machine learning*. McGraw Hill
- Orriols-Puig A, Bernadó-Mansilla E (2006) A further look at UCS classifier system. In *Proceedings of the 8th annual conference on genetic and evolutionary computation workshop program*. ACM Press
- Quinlan JR (1993) *C4.5: Programs for machine learning*. Morgan Kaufmann
- Stone C, Bull L (2003) For real! XCS with continuous-valued inputs. *Evol Comput J* 11(3):279–298
- Wilson S (1995) Classifier fitness based on accuracy. *Evol Comput* 3(2):149–175
- Wilson S (2000a) Get real! XCS with continuous-valued inputs. *Lect Notes Comput Sci* 1813:209–219
- Wilson S (2000b) Mining oblique data with xcs. In *Revised papers of the 3th international workshop on Learning Classifier Systems (IWLCS 2000)*. Springer, pp 158–176

Towards Better than Human Capability in Diagnosing Prostate Cancer Using Infrared Spectroscopic Imaging

Xavier Llorà¹, Rohith Reddy^{2,3}, Brian Matesic², and Rohit Bhargava^{2,3}

¹National Center for Super Computing Applications (NCSA)

²Department of Bioengineering

³Beckman Institute for Advanced Science and Technology

University of Illinois at Urbana-Champaign, Urbana IL 61801

xllora@uiuc.edu, rkreddy2@uiuc.edu, matesic2@uiuc.edu, rxb@uiuc.edu

ABSTRACT

Cancer diagnosis is essentially a human task. Almost universally, the process requires the extraction of tissue (biopsy) and examination of its microstructure by a human. To improve diagnoses based on limited and inconsistent morphologic knowledge, a new approach has recently been proposed that uses molecular spectroscopic imaging to utilize microscopic chemical composition for diagnoses. In contrast to visible imaging, the approach results in very large data sets as each pixel contains the entire molecular vibrational spectroscopy data from all chemical species. Here, we propose data handling and analysis strategies to allow computer-based diagnosis of human prostate cancer by applying a novel genetics-based machine learning technique (NAX). We apply this technique to demonstrate both fast learning and accurate classification that, additionally, scales well with parallelization. Preliminary results demonstrate that this approach can improve current clinical practice in diagnosing prostate cancer.

Categories & Subject Descriptors

I.2.6 [Artificial Intelligence]: Learning—Concept Learning.

I.5.4 [Pattern Recognition]: Applications.

J.3 [Life & Medical Science]: Medical Information Systems.

General Terms

Algorithms, Design, Performance, Experimentation.

Keywords

Genetics-Based Machine Learning, Learning Classifier Systems, Parallelization, Prostate Cancer.

1. INTRODUCTION

Pathologist opinion of structures in stained tissue is the definitive diagnosis for almost all cancers and provides critical input for therapy. In particular, prostate cancer accounts for one-third of noncutaneous cancers diagnosed in US men, and it is a leading cause of cancer-related death. Hence, it is, appropriately, the subject of heightened public awareness and widespread screening. If prostate-specific antigen (PSA) or digital rectal screens are abnormal, a biopsy is considered to detect or rule out cancer. Prostate tissue is extracted, or biopsied, from the patient and examined for structural alterations. The diagnosis procedure involves the removal of cells or tissues, staining them with dyes to provide visual contrast and examination under a microscope by a skilled person (pathologist).

The challenge in prostate cancer research and practice is to provide a novel. Due to personnel, training, natural variability and biologic differences, the challenge in prostate cancer research and practice is to provide accurate, objective and reproducible decisions. Conventional optical microscopy followed by manual recognition has been demonstrated to be inadequate for this task. [18]. Hence, we have recently proposed developing a practical approach to this problem using chemical, rather than morphologic, imaging. [19]. In this approach, Fourier transform infrared imaging (FTIR) is employed to provide the entire vibrational spectroscopic information from every pixel of a sample's microscopy image. While the first steps of developing novel imaging and sampling technologies is now reliable, [7] the computational challenge of providing robust classification algorithms that can rapidly provide decisions remains. Due to the above advances in imaging and sampling, data from thousands of patients is available to train and validate algorithms for different disease states. While the application and type of data are unique, a further confounding factor required efficiently processing large volumes of data generated by FTIR imaging. The classification problem can be formulated as a supervised learning problem in which several million pixels (hundred of gigabytes) of accurately labeled data are available for model training and validation. The volume of tissue and (future) need for intra-operative diagnoses imply that rapid and accurate diagnoses are crucial to allow physicians to explore all possible courses of action. Under these conditions, traditional supervised learning ap-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

GECCO'07, July 7–11, 2007, London, England, United Kingdom.

Copyright 2007 ACM 978-1-59593-697-4/07/0007 ...\$5.00.

proaches and implementations do not scale to provide diagnoses in an appropriate time frame. Hence, efficiently processing and learning models from gigabytes of FTR imaging data requires a careful design of the supervised learning algorithm. Moreover, the biological nature of the problem requires that such models be interpretable to provide fundamental new insight into the disease process. Genetics-based machine learning (GBML) techniques take advantage of the “*quasi embarrassing parallelism*” [17] to provide scaleable, fast, accurate, reliable, and interpretable models. In this paper we present an approach engineered to the desired solution and constraints of addressing this human task. A modified version of a sequential genetics-based rule learner that exploits massive parallelisms via the message passing interface (MPI) and efficient rule-matching using hardware-oriented operations is developed. We named this system **NAX** [24], and we have shown that its performance is comparable to traditional and genetics-based machine learning techniques on an array of publicly available data sets. We now show that **NAX**—taking advantage of both hardware and software parallelism—is able to provide prostate cancer diagnoses that are human-competitive. In this paper, we present preliminary results supporting this outcome.

The paper is structured as follows. Section 2 provides an overview of our approach towards computer-aided diagnoses for prostate cancer. Procedure and form of the data are summarized in section 3. **NAX** is introduced in section 4, where we describe the basic components and design decisions in this approach. In section 5 we present preliminary results indicating that the approach presented in this paper is human-competitive. Finally, section 6 summarizes some conclusions and further research.

2. PROBLEM DESCRIPTION

Prostate cancer is the most common non-skin malignancy in the western world. The American Cancer Society estimated 234,460 new cases of prostate cancer in 2006 [31]. Recognizing the public health implications of this disease, men are actively screened through digital rectal examinations and/or serum prostate specific antigen (PSA) level testing. If these screening tests are suspicious, prostate tissue is extracted, or biopsied, from the patient and examined for structural alterations. Due to imperfect screening technologies and repeated examinations, it is estimated that more than 1 million people undergo biopsies in the US alone.

2.1 Prostate Cancer Diagnosis

The removal of a small section of prostate is most often accomplished by core biopsy. A needle is inserted into the tissue and several (6-23) samples are obtained from different positions. Biopsy, followed by manual examination under a microscope is the primary means to definitively diagnose prostate cancer as well as most internal cancers in the human body. Pathologists are trained to recognize patterns of disease in the architecture of tissue, local structural morphology and alterations in cell size and shape. Specific patterns of specific cell types distinguish cancerous and non-cancerous tissues. Hence, the primary task of the pathologist examining tissue for cancer is to locate foci of the cell of interest and examine them for alterations indicative of disease.

The specific cells in which cancer arises in the prostate

are epithelial cells. While epithelial-origin cancers account for over 85% of all human cancers, they account for more than 95% of prostate cancers. In prostate tissue, epithelial line secretory ducts within the structural cells (collectively termed ‘stroma’) that allow the tissue to maintain its structure and function. Hence, a pathologist will first locate epithelial cells in a biopsy and, to examine for cancer, will mentally segment them from stroma.

Biopsy samples are prepared in a specific manner to aid in recognition of cells and disease. The sample is sliced thin ($\sim 5\mu\text{m}$ thickness), placed on a glass slide and stained with a dye to provide contrast. The most common dye is a mixture of hematoxylin and eosin (*H&E*), which stains protein-rich regions pink and nucleic acid-rich regions blue. Empty space, lipids and carbohydrates are typically not stained and characterized by white color in images. Staining allows the pathologist to identify cells based on their nucleus and extra-nuclear regions. Patterns of the same cell type characterize structures. For example, epithelial cells arranged in a circular manner around empty space are characteristic of a duct and endothelial cells similarly arranged are characteristic of blood vessels. The empty space enclosed within a duct in pathology images is termed a lumen. The distortion of the circular pattern of epithelial cells around a lumen is characteristic of cancer.

In low severity cancers, lumens are only slightly distorted, while higher grades of cancer display a lack of lumen and simply consist of masses of epithelial cells supported by little stroma. The relative distortion and change in lumen shape is organized into a grading scheme to assess the severity of the disease, Gleason Scoring system, which is the primary measure of disease that defines diagnosis, helps direct therapy and helps predict those at danger of dying from the disease. Since prostate cancer is multi-focal and the disease quite variable, two dominant patterns of epithelial distortion are selected and each is independently graded on a scale of 1-5. The grades are then summed to provide a Gleason score ranging from 2 (low grade cancer) to 10 (maximum danger cancer). This scale has been widely used since its creation in the 1960s and currently forms the clinical standard of practice. Manual Gleason scoring, however, has severe limitations.

2.2 Limitations of Current Practice

Widespread screening for prostate cancer has resulted in a large workload of biopsied men [16], placing an increasing demand on services. Operator fatigue is well-documented and guidelines limit the workload and rate of examination of samples by a single operator (examination speed and throughput). Importantly, inter- and intra-pathologist variation complicates decision-making. The consistency in determining Gleason scores is rather poor. Intra-observer measurements show that a pathologist confirms their own score less than 50% of the time and are ± 1 score no more than 80% of cases [2]. Hence, the diagnoses for $\sim 50\%$ of cases may change and may be significantly altered for $\sim 20\%$ of cases ultimately leading to changes in therapy for a patient subset [30]. The numbers are decidedly cause for concern. For example, a recent study including 15 pathologists and 537 prostate cancer patients, 70.8% of Gleason scores were shown to be inaccurate when compared with the patient’s final outcome [18]. Second opinions [29] improve assessment and are cost-effective [10], not to mention their utility in mit-

igating the effects of healthcare costs, lost wages, morbidity, or potential litigation. In summary, the manual recognition of spatial patterns leaves much to be desired from a process perspective and has far-reaching social effects from a public health perspective.

For the reasons underlined above, there is an urgent need for high-throughput, automated and objective pathology tools. We believe that this need is best met by employing the power of computer algorithms and advanced processing to address prostate cancer diagnosis and grading.

The information content of conventionally stained images is limited, inherently non-specific and varies greatly within patient populations and processing conditions. Hence, the information derived from visible microscopy images is fundamentally limited and automated methods of analyzing stained images have failed to provide a sufficiently robust algorithm to diagnose disease. An alternative to morphology-based microscopy are molecular microscopy techniques to probe disease. Molecular technologies for disease diagnosis are an exciting venue for investigations as they promise better diagnostic capabilities through objective means and a multitude of chemicals to provide insight into the changes indicative of the disease process. In particular, spectroscopy tools allow for the measurement of many molecular species simultaneously. Spectroscopic techniques in imaging form, notably using optics, further enable the analysis to be conducted without perturbing the tissue [11]. In this manuscript, we present the analysis of prostate tissue with one such technique, Fourier transform infrared (FTIR) spectroscopic imaging.

2.3 Molecular Imaging

Infrared spectroscopy is a classical technique for measuring the chemical composition of specimens. At specific frequencies, the vibrational modes of molecules are resonant with the frequency of infrared light. By monitoring all frequencies in the region, a pattern of absorption can be created. This pattern, or spectrum, is characteristic of the chemical composition and is hypothesized to contain information that will help determine the cell type and disease state of the tissue. Recently, FTIR spectroscopy has been developed in an imaging sense. Hence, The data are similar to optical microscopy. The first difference is that no external dyes are needed and the contrast in images can be directly obtained from the chemical composition of the tissue. The second is that each pixel in the visible image contains RGB values but in IR imaging contains several thousand values across a bandwidth (2000 – 14000nm) that is ~ 40 times larger than the visible spectrum (400 – 700nm) [7].

3. DATA AND METHODOLOGY

3.1 Experimental Details

Prostate tissues were obtained from Cooperative Human Tissue Network for the tissue array research program (TARP) laboratory. Using these tissues, tissue microarrays were prepared using a Beecher automated tissue arrayer containing a video overlap system and 0.6mm needles. Appropriate institutional review board and National Institutes of Health (USA) guidelines for the protection of human subjects were followed. $5\mu\text{m}$ sections of tissue were floated on an infrared transmissive optical window for FTIR spectroscopic imaging. Another $5\mu\text{m}$ section obtained from the same point

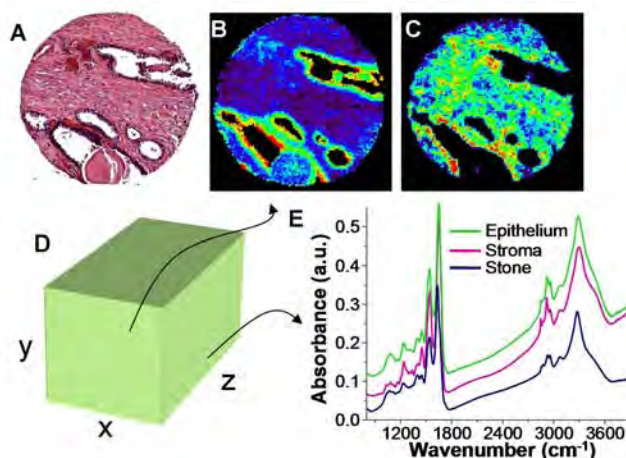


Figure 1: Conventional Staining and Automated Recognition by Chemical Imaging. (A) Typical H&E stained sample, in which structures are deduced from experience by a human. Highlights of specific regions in the manner of H&E is possible using FTIR imaging without stains. (B) Absorption at 1080 cm^{-1} commonly attributed to nucleic acids and (C) to proteins of the stroma. The data obtained is 3 dimensional (D) from which spectra (E) or images at specific spectral features may be plotted.

on the tissue specimen was observed using traditional microscopy for comparison. Expert pathologists determined the tissue classification using these microscopy samples by staining with H&E. Pathologists' classification were used as the 'gold standard' for comparison with the results from the methods mentioned in this paper.

Tissues were analyzed using a Michelson interferometer attached to a microscope (Perkin-Elmer Spotlight 300) in transmission mode at a resolution of 4cm^{-1} . The sample was then raster scanned to obtain images of the entire specimen. Typical specimen size is $600\mu\text{m} \times 600\mu\text{m}$ with each pixel being $6.25\mu\text{m} \times 6.25\mu\text{m}$ on the sample plane. Spectra are composed of 1,641 sample points of the spectral range $4,000 - 720\text{cm}^{-1}$. Data acquisition using these techniques required 40 minutes per cylindrical core of the tissue microarray to yield a root mean square signal to noise ratio of 500 : 1. A typical array was composed of approximately 2.5 million pixels and required 40 GB of storage space.

The data obtained from FTIR imaging is three-dimensional. The x - and y -dimensions locate pixels on the tissue-sample plane. The z -dimension values compose the IR spectrum for the corresponding pixel. The spectra can be analyzed to determine what type of tissue (epithelium, stroma, or muscle) the specimen is as well as whether the tissue is malignant or benign. We have developed this technology to provide data from tissue in minutes and employ a high-throughput sampling strategy using Tissue Microarrays (TMA) to obtain data.[19] Samples from multiple tissues, from multiple patients and multiple clinical settings are included in the data set to maximize the sampling of natural variability and ensure the development of robust analysis algorithms. These high-throughput imaging and

microarray technologies combine to provide very large data sets—see Figure 1. A typical single core consists of 300×300 pixels on the $x - y$ plane with 1641 bands on the z -axis. A tissue microarray consists of several hundred such cores and analysis of such large datasets (typically, tens of GB) is computationally expensive.

3.2 Data Format

Each pixel’s z -dimension contains a spectrum characteristic of the chemical composition of that region of the specimen. Certain spectral quantities provide measures of chemistry. For example, the height of each feature is proportional to its abundance, the peak position is associated with the vibrational identity and peak shape often reflects the multitude of environments around the molecule. Therefore, differences in spectral characteristics can be used in classification and these exact spectral features are termed ‘metrics’. For example, the ratio of absorbance of the spectral peak at 1080cm^{-1} to the spectral peak at 1545cm^{-1} is commonly used to distinguish epithelial from stromal cells. Trained spectroscopists determine these metrics based upon examination of spectral patterns. Hence, the reduction of full spectra to descriptive metrics forms an intelligent dimensionality reduction strategy. Genetic algorithms form decision rules based upon these metrics to classify pixels by tissue type. Furthermore, the transparency of the genetic algorithms allows the scientist to correlate specific rules to biological features (tissue type and cancer classification) via metrics based upon spectral characteristics.

4. APPROACH

In this section we review related work on the GBML community, highlighting previous efforts to deal with large data sets. We also present the motivation and techniques that lead to the design of NAX. Special attention is paid to the description of the hardware and software techniques used, as well as to the design of a scalable GBML algorithm.

4.1 Related Background

Bernadó, Llorà & Garrell [6] presented a first empirical comparison between genetics-based machine learning techniques (GBML) and traditional machine learning approaches. The authors reported that GBML techniques were able to perform as well as traditional techniques. Later on, Bacardit & Butz [3] repeated the analysis again obtaining similar results. Most of the experiments presented on both papers were conducted using publicly available data sets provided by the *University of California at Irvine* repository [28]. Most of the data sets are defined over tens of features and up to few thousands of records. However, a key property of GBML approaches is its intrinsic massive parallelism and scalability properties. Cantú-Paz [8] presented how efficient and accurate genetics algorithms could be assembled, and Llorà [21] presented how such algorithms can be efficiently used as machine learning and data mining techniques.

GBML techniques require evaluating candidate solutions against the original data set matching the candidate solutions (e.g. rules, decision trees, prototypes) against all the instances in the data set. Regardless of the GBML flavor used, Llorà & Sastry [25] showed that as the problem grows, the matching process governs the execution time. For small data sets (teens of attributes and few thousands of records)

the matching process takes more than 85% of the overall execution time marginalizing the contribution of the other genetic operators. This number easily passes 99% when we move to data sets with few hundreds of attributes and few hundred thousands of records. Such results emphasize one unique facet of GBML approaches: scalability via exploiting massive parallelism. More than 99% of the time required is spent on evaluated candidate solutions. Each solution evaluation is independent of each other and, hence, it can be computed in parallel. Moreover, the evaluation process can also be parallelized further on large data sets by splitting and distributing the data across the computational resources. A detailed description of the parallelization alternatives of GBML techniques can be found elsewhere [21].

Currently available off-the-shelf GBML methods and software distributions [5, 20] do not usually target dealing with very large data sets. Three different works need to be mentioned here. Flockhart [12] proposed and implemented GA-MINER, one of the earliest effort to create data mining systems based on GBML systems that scale across symmetric multi-processors and massively parallel multi-processors. The work review different encoding and parallelization schemes and conducted proper scalability studies. Llorà [21] explored how fine-grained parallel genetic algorithms could become efficient models for data mining. Theoretical analysis of performance and scalability were developed and validated with proper simulations. Recently, Llorà & Sastry [25] explored how current hardware can be efficiently used to speed up the required matching of solutions against the data set. These three approaches are the basis of the incremental rule learning proposed in the next section to approach very large data sets—such as the prostate tissue classification one.

4.2 The Road to Tractability

NAX evolves, one at a time, maximally general and maximally accurate rules. Then, the covered instance are removed and another rule is added to the previously stored one, forming a decision list. This process continues until no uncovered instances are left. Llorà, Sastry & Goldberg [26] showed that maximally general and maximally accurate rules [32] could also be evolved using Pittsburgh-style learning classifier systems. Later, Llorà, Sastry & Goldberg [27] showed that competent genetic algorithms [15] evolve such rules quickly, reliably, and accurately. From these early works, it can be inferred that approaching real-world problems, such as the prostate tissue classification and cancer diagnosis, using GBML techniques may produce the desired byproduct: proper scalability. We discuss next efficient implementation techniques to deal with very large data sets using NAX [24].

4.3 Exploiting the Hardware

Recently, multimedia and scientific applications have pushed CPU manufactures to include support for vector instruction sets again in their processors. Both applications areas require heavy calculations based on vector arithmetic. Simple vector operations such as *add* or *product* are repeated over and over. During 80s and 90s supercomputers, such as Cray machines, were able to issue hardware instructions that took care of basic vector operations. A more constrained scheme, however, has made its way into general-purpose processors thanks to the push of multime-

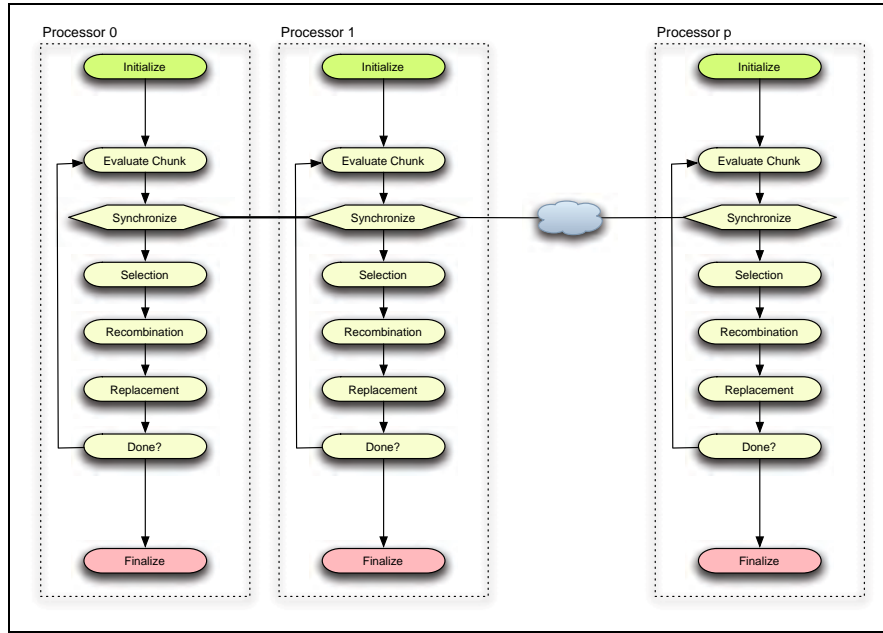


Figure 2: This figure illustrates the parallel model implemented. Each processor is running an identical NAX algorithm. They only differ in the portion of the population being evaluated. The population is treated as collection of chunks where each processor evaluates its own assigned chunk sharing the fitness of these individuals with the rest of processors. This approach minimizes communication cost.

dia and scientific applications. Main chip manufactures—IBM, Intel, and AMD—have introduced vector instruction sets—AltiVec, SSE3, and 3DNow⁺—that allow performing vector operations over packs of 128 bits by hardware. We will focus on a subset of instructions that are able to deal with floating point vectors. This subset of instructions to implemented by hardware vector operations against groups of four floating-point numbers. These instructions are the basis of the fast rule matching mechanism proposed.

Our set of rules seek both to correctly classify the prostate data set and provide biological insight into the rules. All the attributes of the domain are real-value and the conditions of the rules need to be able to express conditions in a \mathbb{R}^n spaces. We use a rule encoding similar to the one proposed by Wilson [33] and widely used in the GBML community. Rules express the conjunction of tests across attributes. Each test can be defined in multiple fashions, but without loss of generality, we pick a simple interval based one. A simple example of and *if-then* rule, could be expressed as follows:

$$1.0 \leq a_0 \leq 2.3 \wedge \dots \wedge 10.0 \leq a_n \leq 23 \rightarrow c_1 \quad (1)$$

Where the condition is the conjunction of the different attribute tests, as introduced earlier, and the condition is the predicting class. We also allow a special condition—**don't care**—which always returns **true** to allow generalized to rules evolve. The rule below illustrates an example of a generalized rule.

$$1.0 \leq a_0 \leq 2.3 \wedge -3.0 \leq a_3 \leq 2 \rightarrow c_1 \quad (2)$$

All attributes except a_0 and a_3 were marked as **don't care**.

Matching a rule requires performing the individual tests before the final *and* condition can be computed. Vector instruction sets can help improve the performance of this process by performing four tests at once. Actually, this process can be regarded as four parallel running pipelines. The

process can be improved further by stopping the matching process when any one test fails. The code implemented assumes that the two vectors containing the upper and lower bounds are provided and records are stored in a two dimensional matrix. As also shown elsewhere [25], exploiting the hardware available can speed between 3 and 3.5 times the matching process[24].

4.4 Massive Parallelism

Since most of the time is spent on the evaluation of candidate rules when dealing with large data sets, our next goal was to find a parallelization model that could take advantage of this feature. Due to the embarrassing parallelism model [17] for rule evaluation, we designed a coarse-grain parallel model for distributing the evaluation load. Cantú-Paz [8] proposed several schemes, showing the importance of the trade off between computation time and time spent communicating. When designing the parallel model, we focused on minimizing the communication cost. Usually, a feasible solution could be a master/slave one—the computation time is much larger than the communication one. However, GBML approaches tend to use rather large populations, forcing us to send rules to the evaluation slaves and collect the resulting fitness. This scheme also increments sequential instructions that cannot be parallelized, reducing the overall speedup of the parallel implementation as a result of Amdahl's law [1].

To minimize communication cost, each processor runs identical NAX algorithms—all seeded in the same manner, and, hence performing the same genetic operations. They only differ in the portion of the population being evaluated. Thus, the population is treated as collection of chunks where each processor evaluates its own assigned chunk, sharing the fitness of the individuals in its chunk with the rest of processors. In this manner fitness can be encapsulated and broadcasted, maximizing the occupation of the underlying pack-

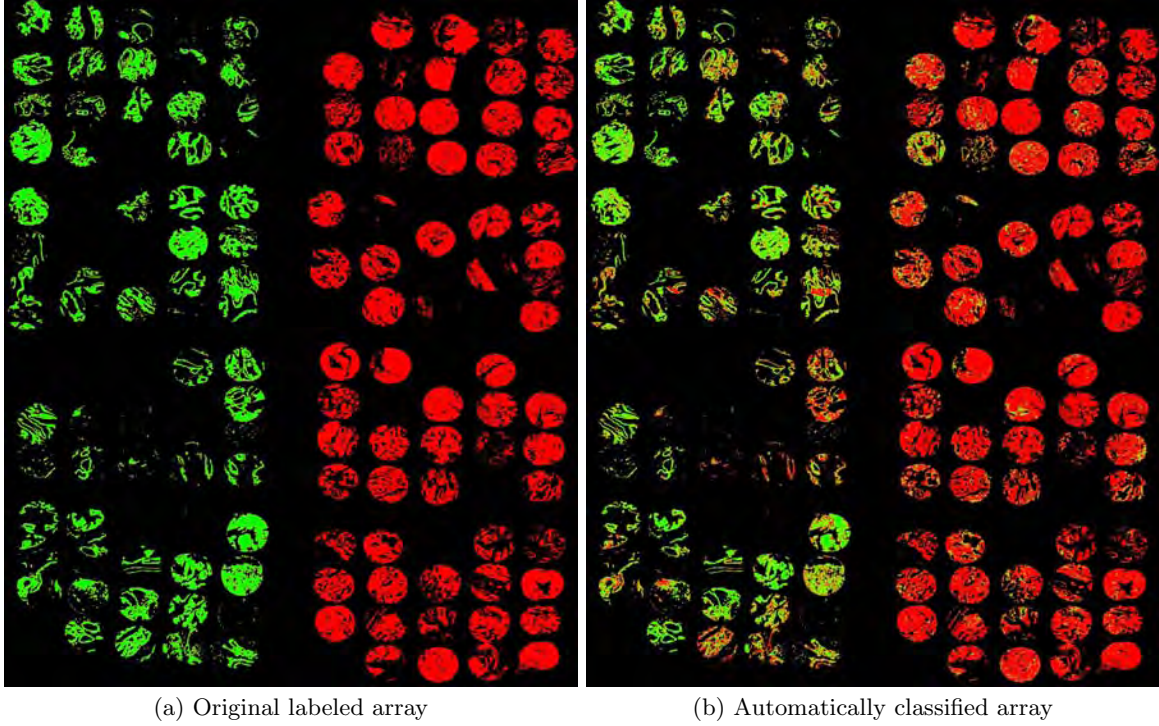


Figure 3: This figure on the left-hand side presents the original labeled data contained in the P80 array. The figure on the right-hand side presents the reconstructed image based on the predictions issued by the rule set evolved by NAX. Green represent non cancerous tissue spots; red represent malignant tissue spots.

ing frames used by the network infrastructure. Moreover, this approach also removes the need for sending the actual rules back and forth between processors—as a master/slave approach would require—thus, maintaining the communication to the bare minimum—namely, the fitness. Figure 2 presents a conceptual scheme of the parallel architecture of NAX.

To implement the model presented in Figure 2, we used C and the *open message passing interface* (openMPI) implementation [13]. Each processor computes which individuals are assigned to it. Then it computes the fitness and, finally, it broadcasts the computed fitness. The rest of the process is unchanged. Except for the cooperative evaluation, all the processors generate the same evolutionary trace.

4.5 Lists of Maximally General and Maximally Accurate Rules

One main characteristic of the so-called Pittsburgh-style learning classifier systems—a particular type of GBML—is that the individuals encode a rule set [14, 22, 15]. Thus evolutionary mechanisms directly recombine one rule set against another one. For classification tasks of moderate complexity, the rule sets are not large. For complex problems, however, the potential number of rules required to ensure accurate classification may use prohibitively large amounts of memory. The requirements increase even further in the presence of noise [23]. Hence, this family of GBML techniques works very well on moderate complexity problems [6, 3], but needs to be modified for complex and large data sets.

A sequential rule learning approach may alleviate the re-

quirements by evolving only one rule at a time, hence, reducing the memory requirements [9, 4]. This allows maintaining relatively small memory footprints that makes feasible processing large data sets. However, an incremental approach to the construction of the rule set requires paying special attention to the way rules are evolved. For each run of the genetic algorithm, we would like to obtain a maximally general and maximally accurate rule, that is, a rule that covers the maximum number of examples without making mistakes [32]. NAX (our proposed incremental rule learner) evolves maximally general and maximally accurate rules by computing the *accuracy* (α) and the *error* (ε) of a rule [26]. In a Pittsburgh-style classifier, the *accuracy* may be computed as the proportion of overall examples correctly classified, and the *error* is the proportion of incorrect classifications issued. Once the *accuracy* and *error* of a rule are known, the fitness can be computed as follows.

$$f(r) = \alpha(r) \cdot \varepsilon(r)^\gamma \quad (3)$$

where γ is the error penalization coefficient. We have set γ to 18 to guarantee that the evolutionary process will produce maximally general and maximally accurate solutions. Further details may be found elsewhere [24]. The above fitness measure favors rules with a good classification accuracy and a low error, or maximally general and maximally accurate rules. By increasing γ , we can bias the search towards correct rules. This is an important element because assembling a rule set based on accurate rules guarantees the overall performance of the assembled rule set. NAX's efficient implementation of the evolutionary process is based on the techniques described using hardware acceleration—section

4.3—and coarse-grain parallelism—section 4.4. The genetic algorithm used was a modified version of the *simple genetic algorithm* [14] using tournament selection ($s = 4$), one point crossover, and mutation based on generating new random boundary elements.

5. RESULTS

NAX has shown competitiveness in evolving rule sets that perform as accurately as the ones evolved by other genetics-based machine learning and non-evolutionary machine learning techniques. However, NAX's key element is the ability to deal with large data sets. In this paper, we present preliminary results towards evolving a model capable of correctly classifying pixels as cancerous or non-cancerous. The original array of spots is presented in figure 3(a). Each spot corresponds to a different biopsy sample from a patient. The pixels present in each spot correspond to the epithelial tissue of the biopsy, we suppress all other tissue types with a prior classification filter based on Bayesian Likelihood.[7] Each pixel of a spot is defined by 93 different metrics extracted from the processed infrared spectra—as described in section 3. Finally, each pixel in the array was labeled with the diagnostic class provided by a human pathologist. Figure 3(a) presents in green all the non-cancerous pixels while red identifies cancerous ones.

Our goal with the initial experiments here was to demonstrate the usefulness of the proposed approach to computer-aided diagnosis. Our current experimental efforts are planning mass experimentation on several tissue arrays using the Tungsten cluster at the National Center for Supercomputing Applications. These initial experiments were conducted on a dual core Intel Xeon 2.8GHz Linux computer with 1Gb of RAM. NAX was run using both processors. The training time to obtain a model describing all the data took less than ten hours—indicating that very competitive training times can be achieved by just using more processors. The obtained model was able to correctly classify $> 99.99\%$ of the training pixels correctly. However, these results do not illustrate the generalization capabilities of the models evolved by NAX. Hence, we ran a series of ten-fold stratified cross-validation runs [34] to measure generalization and test performance of the evolved models. It is important to mention that tools such as WEKA [34] and other off-the-shelf data miners were not able to handle the volume of data required to evolve a model—either due to the large memory footprint required or by not being able to provide an accurate model in a feasible time period. The results of the cross-validation experiments using NAX correctly classified 87.34% of validation pixels. Such results are more than encouraging, because they show a human-competitive computer-aided diagnosis system is possible. Another interesting property is that a few rules classify a large number of pixels—see Figure 4. Such a result is interesting for the interpretability of the model, since a small number of rules have a great expressiveness, and hence may provide valuable biological insight. Most importantly, they allow us to classify tissue accurately. Subsequent to this pixel level classification, each circular spot in figure 3 was assigned as malignant or benign based on the majority of pixels of the class in the sample. We were able to accurately classify 68 of 69 malignant spots and 70 of 71 benign spots in this manner. While human accuracy is difficult to quantify due to the variation between persons, a generally accepted anecdotal figure is about 5%

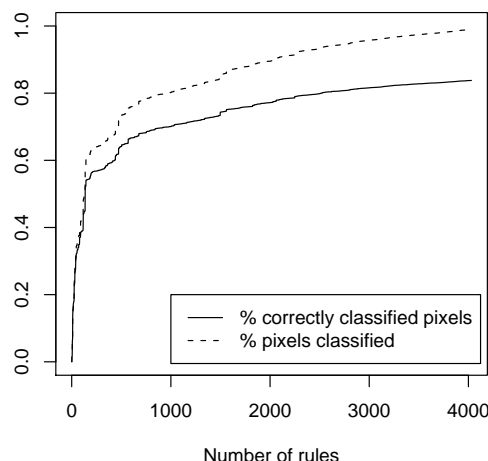


Figure 4: Performance of the evolved model as a function of the number of rules used.

error rates. The preliminary results we demonstrate here could potentially reduce that five-fold to about 1%, providing a solution to this real-world problem by a combination of novel spectroscopy and advanced machine learning.

6. CONCLUSION

In this manuscript, we present the application of advanced genetics-based machine learning algorithms to a real-world problem of large scope, namely, the diagnosis of prostate cancer. As opposed to subjective human recognition of disease in tissue using light microscopy, we employed a chemical microscopy approach that required extensive computation but provided a decision without human input. Our development of a learning algorithm based on maximally general and maximally accurate rules was scalable to very large data sets and parallelized to provide learning and classification speed advantages. The algorithm was able to classify a majority of pixels correctly, resulting in overall error rates that were comparable to human examination, the current gold standard of care.

7. ACKNOWLEDGMENTS

This work was sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant FA9550-06-1-0370, the National Science Foundation under grant ISS-02-09199 at the National Center for Supercomputing Applications, UIUC. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation thereon.

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research, the National Science Foundation, or the U.S. Government. We would like to thank Dr. Stephen M. Hewitt, MD PhD, who provided us the samples and expert pathol-

ogy consultation. The research is also supported by grants to RB from the Research Board, UIUC; the Department of Defense through the Prostate Cancer Research Program and by the National Center for Supercomputing Applications, under the auspices of the NCSA/UIUC faculty fellows program.

8. REFERENCES

- [1] G. Amdahl. Validity of the single processor approach to achieving large-scale computing capabilities. In *Proceedings of the American Federation of Information Processing Societies Conference (AFIPS)*, volume 30, pages 483–485. AFIPS, 1967.
- [2] M. Amin, D. Grignon, P. Humphrey, and J. Srigley. *Gleason Grading of Prostate Cancer: A Contemporary Approach*. Lippincott Williams & Wilkins: Philadelphia, 2004.
- [3] J. Bacardit and M. Butz. *Advances at the frontier of Learning Classifier Systems (Volume I)*, chapter Data Mining in Learning Classifier Systems: Comparing XCS with GAssist, page in press. Springer-Verlag, 2006.
- [4] J. Bacardit and N. Krasnogor. Biohel: Bioinformatics-oriented hierarchical evolutionary learning. Nottingham eprints, University of Nottingham, 2006.
- [5] A. Barry and J. Drugowitsch. LCSWeb: the LCS wiki, 1997. <http://lcsweb.cs.bath.ac.uk/>.
- [6] E. Bernadó, X. Llorà, and J. Garrell. *Advances in Learning Classifier Systems: 4th International Workshop (IWLCS 2001)*, chapter XCS and GALE: a Comparative Study of Two Learning Classifier Systems with Six Other Learning Algorithms on Classification Tasks, pages 115–132. Springer Berlin / Heidelberg, July 2001.
- [7] R. Bhargava, D. Fernandez, S. Hewitt, and I. Levin. High throughput assessment of cells and tissues: Bayesian classification of spectral metrics from infrared vibrational spectroscopic imaging data. *Biochemica et Biophysica Acta*, pages 830–845, 2006.
- [8] E. Cantú-Paz. *Efficient and Accurate Parallel Genetic Algorithms*. Kluwer Academic Publishers, 2000.
- [9] O. Cordón, F. Herrera, F. Hoffmann, and L. Magdalena. *Genetic Fuzzy Systems. Evolutionary tuning and learning of fuzzy knowledge bases*. World Scientific, 2001.
- [10] J. Epstein, P. Walsh, and F. Sanfilippo. Clinical and Cost Impact of Second-opinion Pathology: Review of Prostate Biopsies Prior to Radical Prostatectomy. *American Journal of Surgical Pathology*, 20:851–857, 1996.
- [11] D. Fernandez, R. Bhargava, S. Hewitt, and I. Levin. Infrared spectroscopic imaging for histopathologic recognition. *Nature Biotechnology*, 23(4):469–474, 2005.
- [12] I. Flockhart. GA-MINER: parallel data mining with hierarchical genetic algorithms (final report). Technical Report Technical Report EPCCA/KMS-GA-MINER-REPORT 1.0, University of Edinburgh, 1995.
- [13] E. Gabriel, G. Fagg, G. Bosilca, T. Angskun, J. Dongarra, J. Squyres, V. Sahay, P. Kambadur, B. Barrett, A. Lumsdaine, R. Castain, D. Daniel, R. Graham, and T. Woodall. Open MPI: Goals, concept, and design of a next generation MPI implementation. In *Proceedings of the 11th European PVM/MPI Users' Group Meeting*. Springer, 2004.
- [14] D. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional., 1989.
- [15] D. Goldberg. *The Design of Innovation: Lessons from and for Competent Genetic Algorithms*. Springer, 2002.
- [16] S. J. Jacobsen, S. K. Katusic, E. J. Bergstralh, J. E. Oesterling, O. Del, G. G. Klee, C. G. Chute, and M. M. Lieber. Incidence of prostate cancer diagnosis in the eras before and after serum prostate-specific antigen testing. *JAMA*, 274:1445–1449, 1995.
- [17] V. Kumar, A. Grama, A. Gupta, and G. Karpis. *Introduction to Parallel Computing: Design and Analysis of Parallel Algorithms*. Benjamin-Cummings Publishing Company, 1994.
- [18] J.-B. Lattouf and F. Saad. Gleason score on biopsy: is it reliable for predicting the final grade on pathology? *BJU International*, 90:694–699, 2002.
- [19] I. Levin and R. Bhargava. Fourier transform infrared vibrational spectroscopic imaging: integrating microscopy and molecular recognition. *Annual Review of Physical Chemistry*, 56:429–474, 2005.
- [20] X. Llorà. Learning Classifier Systems and other Genetics-Based Machine Learning Blog, 2006. <http://www-illigal.ge.uiuc.edu/lcs-n-gbml/>.
- [21] X. Llorà. *Genetics-Based Machine Learning using Fine-grained Parallelism for Data Mining*. PhD thesis, Enginyeria i Arquitectura La Salle. Ramon Llull University, Barcelona, Catalonia, European Union, February, 2002.
- [22] X. Llorà and J. Garrell. Knowledge-independent data mining with fine-grained parallel evolutionary algorithms. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO'2001)*, pages 461–468. Morgan Kaufmann Publishers, 2001.
- [23] X. Llorà and D. Goldberg. Bounding the effect of noise in multiobjective learning classifier systems. *Evolutionary Computation Journal*, 11(3):279–298, 2003.
- [24] X. Llorà, A. Priya, and R. Bhargava. Observer-invariant histopathology using genetics-based machine learning. Technical report, Illinois Genetic Algorithms Laboratory, University of Illinois at Urbana-Champaign (IlligAL TR No 200627), 2006.
- [25] X. Llorà and K. Sastry. Fast rule matching for learning classifier systems via vector instructions. In *Proceedings of the 2006 Genetic and Evolutionary Computation Conference*, pages 1513–1520. ACM Press, 2006.
- [26] X. Llorà, K. Sastry, and D. Goldberg. The compact classifier system: Motivation, analysis and first results. In *Proceedings of the Congress on Evolutionary Computation*, volume 1, pages 596–603. IEEE press, 2005. (Also as IlligAL TR No. 2005019).
- [27] X. Llorà, K. Sastry, D. Goldberg, and L. de la Ossa. The χ -ary extended compact classifier system: Linkage learning in Pittsburgh LCS. In *Advances at the frontier of Learning Classifier Systems (Volume II)*, page in preparation. Springer, 2007. IlligAL report no. 2006015.
- [28] C. J. Merz and P. M. Murphy. UCI repository for machine learning data-bases, 1998. <http://www.ics.uci.edu/~mllearn/MLRepository.html>.
- [29] W. Murphy, I. Rivera-Ramirez, L. Luciani, and Z. Wajzman. Second opinion of anatomical pathology: A complex issue not easily reduced to matters of right and wrong. *J. Urol*, 165:1957–1959, 2001.
- [30] J. Nguyen, D. Schultz, A. Renshaw, R. Vollmer, W. Welch, K. Cote, and A. D'Amico. The impact of pathology review on treatment recommendations for patients with adenocarcinoma of the prostate. *Urologic Oncology: Seminars and Original Investigations*, 22:295–299, 2004.
- [31] A. C. Society. How Many Men Get Prostate Cancer?, 2006. http://www.cancer.org/docroot/CRI/content/CRI_2_2_1X.How_many_men_get_prostate_cancer_36.asp?rnav=cri.
- [32] S. Wilson. Classifier fitness based on accuracy. *Evolutionary Computation*, 3(2):149–175, 1995.
- [33] S. Wilson. Get real! XCS with continuous-valued inputs. *Lecture Notes in Computer Science*, 1813:209–219, 2000.
- [34] I. H. Witten and E. Frank. *Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, CA., 2000.

Fourier transform infrared spectroscopic imaging: the emerging evolution from a microscopy tool to a cancer imaging modality

Gokulakrishnan Srinivasan and Rohit Bhargava

*Department of Bioengineering and Beckman Institute for Advanced Science and Technology,
University of Illinois at Urbana-Champaign, Urbana, IL 61801*

INTRODUCTION

The integration of FTIR spectroscopy with microscopy facilitates recording of spatially resolved spectral information, allowing the examination of both the structure and chemical composition of a heterogeneous material. While the first such attempt was over 50 years ago,¹ present day instrumentation largely evolved from the point microscopy detection of interferometric signals that developed in the mid-80s.² The successful coupling of interferometry for spectral recording and microscopy for spatial specificity in these systems spurred interest in a variety of fields, including the materials,³ forensic⁴ and biomedical arenas.^{5,6} Point microscopy utilizes an aperture to restrict radiation incident on a sample and permits the recording of spatially localized data. The primary utilities of this form of microscopy lay in acquiring accurate spectra from small-size samples, in determining the chemical structure and composition of the heterogeneous phases at specified points and in building a two-dimensional map of the chemical composition of samples. Since the data were acquired at a single point, composition maps could only be acquired by rastering the sample. Hence, the approach was termed mapping or point mapping and involved as many spectral scans as the number of pixels in the map.

The use of focal plane array (FPA) detectors for microscopy^{7,8} allowed for the acquisition of large fields of view in a single interferogram acquisition sweep. The multichannel detection enabled by array detectors was similar to the concept of recording images with charge coupled devices in optical microscopy; hence, the approach was termed imaging. The unique advantages of observing an entire field of view rapidly permitted applications that allowed monitoring of dynamic processes, spatially resolved spectroscopy of large samples or many samples and enhancement of spatial resolution due to retention of radiation throughput that was lost in point microscopy systems due to diffraction at the aperture. Just as for the previous generation of microspectroscopy instruments, applications rapidly followed in the materials⁹ and biomedical fields.¹⁰⁻¹⁴ Research activity in this area can be divided into three major categories: instrumentation and sampling methodologies, applications and data extraction algorithms. In this manuscript, we review key advances and recent developments in the context of biomedical imaging. We do not provide comprehensive overview but selectively highlight certain features of importance for cancer-related imaging. Last, we focus on one emerging application area, namely tissue histopathology, and provide illustrative examples from our laboratory in indicating the integrative nature of the three in developing protocols.

INSTRUMENTATION, SAMPLING AND DATA HANDLING TECHNIQUES

Instrumentation

Since imaging is largely based on new detectors with unique performance characteristics for spectroscopy, efforts in instrumentation have largely focused on the efficient integration of FPA detectors with interferometers. Due to the size, different electronics and unique noise characteristics of FPAs, an optimization of data acquisition methodology was a primary activity in the initial time period of availability of instrumentation. The first rational attempt at understanding performance and optimizing the data acquisition process revealed the unique noise characteristics that limited the first generation of array detectors.¹⁵ Briefly, this paper established that the general behavior of FTIR spectrometers is generally held for imaging spectrometers but the detector may serve to limit the applicability of established practices in IR spectroscopy. An explicit optimization of the data acquisition time revealed several strategies for speeding data collection for both the step scan and rapid scan mode.¹⁶ The first example of rapid-scan FTIR imaging¹⁷ was conducted using asynchronous sampling, followed by descriptions of synchronously triggered sampling and generalized methodologies¹⁸ that could use any detector at any modulation frequency using post-acquisition techniques. Advances in detector technology have now allowed for rapid scan imaging to become routine for large FPA detectors, while innovative new detectors have been developed (first by PerkinElmer) that trade off a large multichannel detection advantage of arrays against the speed of smaller detector arrays to provide a very high performance instrument.¹⁹

At present, rapid scan imaging has become the mode of choice for most manufacturers and detector sizes have proliferated from the classic 64 x 64 format to range from 16 x 1 to 256 x 256 formats (see figure 1). While the smaller detectors require rastering to image most samples and can provide data of higher quality more efficiently, larger detectors are generally employed for their large field of view and are useful for studying dynamics. It is interesting to note that the linear array approach has an entirely different detector technology and considerations for electronics compared to the two-dimensional FPAs. While it is beyond the scope of this article to discuss the differences, the use of “macro” electronics that are offset from the actual detector and AC mode of operation are the two major differences that affect data. Consequently, comparisons in performance are slightly more complicated. On the large format FPA front, the latest advance seems to be a detector developed jointly by NIH and FBI personnel in 2005. The detector can operate at 16 KHz for 128 x 128 pixel snaps (Bhargava, Levin, Perlman and Bartick, *Unpublished*). This is in the speed regime of single element detectors. Hence, the development can truly lead to the acquisition of an entire image in a single interferometer mirror sweep in the same time that it takes to acquire 1 spectrum with a benchtop IR spectrometer. To handle the large data output, we designed on-chip co-addition and various corrections. We believe that similar detector systems, operating in a fast regime and integrating processing with electronics, are likely to be the technology of tomorrow for FTIR imaging.

The wide variety of instrumentation makes comparisons difficult, especially when manufacturers provide different specifications for instruments. We have proposed a comparison index for these systems based on performance per unit time. Recognizing that spectral resolution, time for scanning, data processing (e.g. apodization) and resultant image size are the primary determinants of performance, a measure can be formulated to describe performance. For a fixed

data processing scheme (filtering, apodization etc.), the time taken to acquire 1 megapixels of data for 8 cm^{-1} resolution at a signal to noise ratio (SNR) of 1000:1 is found to be a good measure. We would like to emphasize that the performance is the performance of the entire imaging spectrometer and not due to the detector alone. Efficient coupling of the interferometer and optimization of the optical train will both affect performance as will the correct setup of the experiment. This index also does not consider the ease of use or “user-friendliness” of systems. These are other important considerations and must also be considered by organizations interested in FTIR imaging technology. The issue of time resolution for acquiring data is one such concern. The first approach is the kinetics approach in which the interferometer is repeatedly scanned and imaging data sets are sequentially acquired as quickly as possible. Clearly, rapid scan is favored and the availability of fast readout detectors is mandatory for fast events. The limit to this method is the readout speed of the array (frames in ms) as interferometers can generally be scanned fast enough and the integration time required is typically in the tens of microseconds regime. An example is shown in figure 2 to demonstrate applicability in monitoring polymerization kinetics.

Though rapid scan imaging has displaced the step-scan mode in most new instrumentation, a very important application of the step-scan approach remains in time-resolved imaging.²⁰⁻²² Briefly, the method is applicable to systems that can be repeatedly and reproducibly excited and relax back to their ground state. At each mirror retardation, the FPA is repeatedly triggered to acquire data. At the same time, the sample is excited once and the dynamics of excitation and decay of the excited state are monitored. Mirror stepping, data acquisition and sample excitation are all precisely synchronized. Figure 3 demonstrates the synchronization. Time resolved FTIR imaging was first demonstrated using polymer-liquid crystal composites. Examples of the types of data that may be obtained are also shown in figure 3. Last, the technology was extended to provide significantly higher time resolution than could be obtained by the electronics of the detector alone.²³ While FPA detectors are slow compared to single point detectors used in conventional FTIR spectroscopy, the cause is the need to read out data from several thousand pixels and not from the need to record data from all pixels. Hence, by staggering the data recording time over multiple sample excitations, higher temporal resolution may be obtained. With current detectors, a time resolution of $\sim 30\text{ }\mu\text{s}$ should be possible.

Sampling

Interferometer Issues

Among the sampling configurations, the first clearly was the optimization of the microscope for transmission and sampling. Unexpected issues were encountered in initial devices. For example, the detector for the mono-wavelength laser provides a fringe pattern to allow for tracking mirror retardation. The signal from this laser is measured by a small detector located at the center of the beamsplitter (to minimize errors) with an arm that extends out to the edge. When imaged onto the FPA, this laser detector leads to a pattern with low signal levels. Hence, the field of view is not uniform, leading in turn, to lower signal to noise ratios (SNR) for the affected region. Many manufacturers, hence, have re-designed their spectrometers for imaging use. Another manufacturer has avoided this issue by aligning their microscope to sample only the unaffected part of the beam. Since the non-imaging spectrometer did not require imaging and the interferometer was simply coupled to a microscope, these issues were slowly addressed.

Sampling Modes: Transmission, Transmission-reflection, Reflection and Attenuated Total Reflection

A vast majority of studies report the use of transmission sampling. Other major developments have been the incorporation of reflective slides,^{24, 25, 26} the integration of ATR elements for both microscopy and large sample imaging, integration of ATR technology with various sample forming accessories, grazing angle accessories and multi-sample accessories. Reflective slides actually result in reflection-absorption that allows the beam to sample the signal twice, though with a different phase and lower signal due to half the objective being used for transmitting light to the sample and the other half being used to acquire light from it. A detailed theoretical understanding of the confounding effects has not been published, though an example of the possible data correction algorithm has been reported. ATR imaging is also highly prevalent and available as attachments to conventional imaging microscopes, using the sample chamber of the spectrometer and using it as a solid immersion lens.²⁷ We discuss examples of ATR imaging next.

ATR

In the Attenuated Total Reflection (ATR) mode, an IR transmitting crystal of precise geometry of high refractive index is employed as a solid immersion lens. Light is totally reflected at the sample-crystal interface and an evanescent field penetrates into the sample to provide the interaction to be observed using the traveling wave. Since the sample interaction is largely determined by the lens and not by the sample, precise and controlled depth of interaction is available. The sample, however, needs to be in good contact to allow efficient coupling with the evanescent wave. ATR imaging allows users to work with relatively thick sample sections that do not require much sample preparation expertise or time. The first use of ATR imaging was reported by Digilab in analyzing large samples that were not sectioned, as for transmission. ATR imaging microscopy was demonstrated soon after,²⁸ followed by other novel accessories. There were other unpublished attempts that one of the authors is aware of: In 1999, for example, Snively *et al.* (personal communication, unpublished) demonstrated imaging data from an inverted ZnSe prism acting as a single bounce ATR. Soon after, we employed a Ge crystal but found the signal to noise ratio of the imaging system of that time to be very poor. In addition to the ease of sample preparation, another major advantage of ATR imaging lies in improving the limited spatial resolution of transmission microscopy.²⁹ The authors assessed that they were able to achieve a spatial resolution of 1 μm with a Ge internal reflection element.

Both micro and macro sampling has been extensively utilized.³⁰ A spatial resolution of 3-4 μm using a Ge ATR element was claimed based on more stringent criteria than used previously.²⁹ Ge, ZnSe and diamond³⁰ crystals have been the materials of choice for most applications. In particular, Kazarian and co-workers have extensively employed ATR-FTIR imaging for various applications including drug release; polymer/drug formulations and biological systems.³⁰⁻³³ The same group has provided other innovative sampling configurations for specific experiments, including a compaction cell that allows compaction of a tablet directly on a diamond crystal with a subsequent imaging.³⁴ The changes in the distribution of a tablet consisting of hydroxypropyl methylcellulose (HPMC) and caffeine upon contact with water were studied. In this manner, conventional dissolution measurements were combined with a concurrent assessment of the compacted tablet structure.³⁵ As opposed to the organic solvent-polymer dissolution experiments reported earlier, this configuration allows for easy handling and imaging of water-induced dissolution. The setup can also provide high throughput analysis of materials under controlled

environments.³⁶ Microdroplet sample deposition system was combined with a humidity control device to image about 100 samples deposited on the surface of an ATR crystal simultaneously. The approach was extended to 165 samples and we reported to study parallel dissolution of formulations.³⁷

Multi-sample Accessories and Sampling

While imaging the structure of materials has been the primary focus of FTIR imaging, a number of applications utilize the imaging of multiple samples. The first examples were from the field of catalyst research.³⁸ Typically 2-12 samples could be imaged and analyzed under the same conditions. High throughput validation or method development was the primary goal in these studies. Tissue microarrays (TMAs) provide the same function in biomedical imaging. TMAs consist of tens to hundreds of samples arranged on a grid format. This allows for easy visualization of the structure and classification accuracy across many patients and the statistical measures needed for rigorous validation. The primary utility of the multisample image in this case is to provide wide-ranging sampling and convenient archiving or data storage, not necessarily to provide a higher throughput.^{14,39} With the appropriate geometry, many samples can be imaged to understand their dynamics in a concerted fashion. To accommodate the samples, the field of view is often expanded. This results in a lower spatial resolution. For imaging multiple samples, though, the spatial resolution can be conserved but temporal resolution is restricted.

BIOMEDICAL APPLICATIONS

Bone

Bone has been the tissue studied most by FTIR imaging. Bone composition changes with development, environment, genetics, health and disease, is amenable to imaging at the resolution length scale of imaging and has a limited chemical composition that is characterized using IR spectroscopy.⁴⁰ For almost 30 years until the late 1980s,⁴¹ bone structure was studied using single element detectors in FTIR spectrometers. Typically, ground bone was analyzed using the conventional KBr pellet method. This pellet method obviously destroyed local structures, precluding an understanding of molecular variations due to disease. Nevertheless, it was sensitive to chemical composition and did provide useful information. With microscopy and now with FTIR imaging, sample integrity is maintained and ability to acquire spectral information at anatomically discrete sites is possible. From the resulting spectra, several important pieces of information can be obtained. For example, a) relative mixture composition of hydroxyapatite and collagen by calculating the ratio of the integrated ν_1 , ν_3 phosphate and amide I (mineral: matrix ratio), b) carbonate substitution by calculating the ratio of carbonate/phosphate ratio from the ratio of integrated ν_2 carbonate peak ($850\text{-}900\text{ cm}^{-1}$) and ν_1 , ν_3 phosphate contour ($900\text{-}1200\text{ cm}^{-1}$), c) crystallinity of the mineral phase from the ratio of 1030/1020 peak intensity.⁴² These assays illustrate several quantities important to bone research and disease diagnoses that can be readily performed. Though a complete discussion is available in the reference^{40,42-44}, we pick three illustrative examples demonstrating the applicability in disease and in research.

IR spectral analysis of healthy and diseased bone has been reviewed by Boskey *et al.*⁴² with particular emphasis on changes in bone composition, physiochemical status of mineral and matrix of bones during osteoporosis and the effect of therapeutics on these parameters. Osteoporosis or porous bone is a bone disease characterized by low bone mass and structural

deterioration of bone tissue. This leads to bone fragility and an increased susceptibility to fractures, especially at the hip, spine and wrist. FTIR images of the mineral content and crystallinity in trabecular bone of normal and osteoporotic samples clearly depicts that the trabeculae in diseased tissue are thinner. Moreover, the mineral/matrix ratio in osteoporotic bone is significantly reduced, whereas crystallinity is increased. These advances demonstrate the potential and applicability of the technique to characterize diseased tissue. Bone mineral changes between a healthy mouse model and Fabry diseased (lipid storage disease) mouse model were also analyzed in which globotriaosyl ceramide (Gb3) accumulates in tissues.⁴³ No significant differences in the bone mineral properties were observed between Fabry and healthy mice, which might reflect the similar lack of major bone phenotype in human patients with Fabry's disease and may also be related to the developmental age of the animals. The study provides an example of the applicability to laboratory research.

Calcified tissue in biopsies from adults with osteomalacia has been studied.⁴⁴ Osteomalacia results in a deficiency of the primary mineralization of the matrix, leading to an accumulation of osteoid tissue and reduction in bone's mechanical strength. A decrease in trabecular bone content with absence of changes in matrix or mineral is noticed when iliac crest biopsies of individuals with vitamin D deficient osteomalacia are compared to normal controls. These findings support the assumption that, in osteomalacia, the quality of the organic matrix and of mineral in the centre of the bone does not vary, whereas less-than optimal mineralization occurs at the bone surface.

Brain

Monkey brain tissues were one among the first tissues examined by using FTIR imaging.¹² Lately, the applications have experienced a renaissance with applications to the human brain. Grossly, brain can be divided into two types of matter, namely gray matter and white matter. These names derive simply from their appearance to the naked eye. Gray matter consists of cell bodies of nerve cells while white matter consists of the long filaments that extend from the cell bodies - the "telephone wires" of the neuronal network, transmitting the electrical signals that carry the messages between neurons. A visualization of the two compartments formed the first demonstrative application of FTIR microspectroscopic imaging.

FTIR imaging and multivariate statistical analyses (unsupervised hierarchical cluster analysis) were applied along with histology and immunohistochemistry in an animal model having Glioblastoma multiform (GBM).⁴⁵ GBM is a highly malignant human brain tumor that is considered to be the one of the most difficult to treat effectively.⁴⁶ Authors were able to identify the tumor growth as chemically distinct from the surrounding brain tissue. The distribution of the absorbance of amide I in images highlighted high concentrations of proteins in the corpus callosum and regions of basal ganglia for healthy brain. Low absorbance was generally observed in the cortex, whilst a higher absorbance was observed at outer layer of the cortex. For a GBM bearing animal, the highest absorbance was found at the tumor site. In contrast to healthy brain, a lower absorbance of the amide I band was observed at the corpus callosum when compared to that in the cortex and the caudoputamen. The study demonstrates a powerful application of simple analyses that can indicate disease. It also highlights the multitude of spatial and spectral clues that can be used to diagnose or understand the disease.

In addition to primary disease sites, diagnoses metastatic spread from various cancers was also reported.⁴⁷ A multivariate classification algorithm was used to distinguish normal tissue from brain metastases successfully and to classify the primary tumor of brain metastases from renal cell carcinoma, lung cancer, colorectal cancer, and breast cancer. In the cluster-averaged IR spectra from a brain metastasis of renal cell carcinoma, the main spectral differences were observed for the three tissue regions in the region from 950 to 1200 cm^{-1} and from 1500 to 1700 cm^{-1} . Band intensities of 1026, 1080 and 1153 cm^{-1} are at maximum in the spectrum of black cluster and minimum in the spectrum of light gray cluster. The comparisons of the IR spectra of normal brain tissue and brain metastases of lung, breast cancer and colorectal cancer were made and found that these spectra do not contain spectral features at 1026, 1080 and 1153 cm^{-1} that are indicative of the presence of glycogen. It was concluded that these aforementioned spectral features would be considered as biomarkers for brain metastases of the primary tumor renal cell carcinoma. In addition to these three bands, the spectral differences were observed for the bands at 1542 and 1655 cm^{-1} , owing to the presence of amide I and amide II vibrations. It is clear from the results that the maximum protein concentrations correlate with minimum glycogen concentrations in the IR image. However, the protein and glycogen properties evident in the IR image are not visible in the unstained cryosection. It is noteworthy that simple univariate analyses provide the end clues to the disease. Even on application of multivariate techniques, the most prominent and easy to understand biomarkers of disease are those defined by conventional spectroscopic knowledge as being important for identification, namely, features and their absorption.

In the cluster-averaged IR spectra of white matter from the three normal brain tissue samples, intense bands at 1060, 1233, 1466, 1735, 2850 and 2920 cm^{-1} due to the high lipid concentration in white matter were noticed. Intensity changes were due to inter-sample and patient to patient variances of the same tissue type. In addition, cluster-averaged IR spectra of a brain metastasis of (renal cell carcinoma, breast cancer, lung cancer, and colorectal cancer) and gray matter of normal brain tissue were compared after baseline subtraction and then normalization with respect to the amide I band. Significant differences in the band positions, intensities and area were observed between these samples which were then used as potential candidates to differentiate normal and tumor tissue and for the identification of the primary tumor. Here, authors used only eight spectroscopic features for LDA model. They were able to classify correctly for three out of three normal brain tissue and 16 out of 17 brain metastases samples. Hence, though univariate analyses and features provide useful recognition, their integration into a multivariate algorithm provides for automated recognition of clinical importance. It may also be argued, however, that it is questionable whether the small numbers of samples employed represent a true performance condition for the algorithm or are simply reflective of bias arising from the clinical setting or sample sources. The advent of faster imaging approaches and advanced sampling techniques like TMAs can allow for larger numbers of samples to be analyzed and such doubts about the validity of studies be put to rest.

Similarly, tissues from rat Glioma models have been characterized and used to discriminate healthy from tumor sections using principal component analysis and K-means.⁴⁸ Pseudo color maps reported were constructed on 8-means clusters, where each cluster is consisting of similar spectra. The lipids/protein ratio (1466/1452 cm^{-1}) was found to be decreased and the band at 1740 cm^{-1} became weak and almost vanished as compared to the corresponding bands in the

healthy tissue. In addition to the above mentioned differences, significant differences between healthy and tumor affected tissue were observed in the finger print region. In the healthy tissue, a weak band at 1172 cm^{-1} , representing the stretching mode of C-O groups were observed. Reduced intensity as well as shifting of peak to 1190 cm^{-1} was noted for tumor and surrounding tumor spectra. Tumor tissue was observed to contain a decreased intensity of the asymmetric phosphate stretching and C-C stretching and an increased intensity of the symmetric phosphate stretching when compared to the healthy tissue. Variations in lipid features (methylene and methyl stretching) were also observed. The major point here is that the entire spectrum contains numerous points of difference between healthy and diseased tissue. Results were found to be in agreement with those obtained from pathology.⁴⁹ The structural difference around the tumor was noted, which could be ascribed to the peritumoral edema observed during glioma development. An increase in the permeability of the blood-brain barrier and aggravation in the mass effect of tumors are the rationale for edema, which is associated with brain tumor. Fundamental understanding can be enhanced by a complete understanding of the spectral differences but prediction algorithms need only a few measures of the spectral data to be effective.

Breast

Two major applications in breast tissue deal with complications arising from artificial alterations of the tissue and the evolution of cancer. While breast augmentation by implants is highly prevalent, its complications have been discussed more recently. On the other hand, the conventional method for diagnosing and evaluating the prediction of breast disease is a histopathological examination of biopsy samples, a practice that has some shortcomings. For breast implants, a major question is the containment of filling material as its leakage can lead to potential diseases. The silicone gel in implants is very different chemically from surrounding tissue and its presence in tissue sections indicates a definite leak from the implant either due to material failure as a consequence of aging. A spectroscopic image⁵⁰ generated from the asymmetric stretching modes of the methyl groups attached to silicon in the gel allowed for the examination of silicone in the tissue. Due to the unique chemical contrast employed in FTIR imaging, such presence can be discerned within the tissue, even when optical microscopy contrast was poor. An example of presence of Dacron (a commercial name for poly(ethylene terephthalate)) fixative patch threads in the breast tissues was shown.⁵⁰ It was noted that the technique is capable of rapid analysis within minutes of sectioning the tissue.

A few reports have also applied FTIR imaging for diagnosing breast diseases. Breast tumor tissues were characterized by both FTIR imaging and point mapping techniques and advantages over the other were evaluated.⁵¹ Similar comparisons had previously been reported for polymeric materials, analyzing both static and dynamic samples.⁵² Comparison images from the two methods, imaging data provided a clearer structure in the tumor area than the data obtained from point mapping. Since breast tumor cells are $\sim 10\text{ }\mu\text{m}$ in diameter, point mapping data (with an aperture of $30\text{ }\mu\text{m}$) would always contain the spectrum of tumor cells as well as from the contributions of other components surrounding the cells. The study clearly indicated that the conventional point mapping approach can fail to detect a small number of malignant cells due to its poor resolution capabilities. Nevertheless, the contamination problem, i.e., the spectral contributions of other components surrounding the cell is found to be less severe in case of ductal carcinoma in situ (DCIS). The study illustrates the need for matching the appropriate level

of spatial resolution to the task. While the 30 μm resolution may be appropriate for some applications, it was clearly insufficient for detecting smaller numbers of cells.

Artificial network and K-means cluster analysis have also been employed for the classification of FTIR imaging data from normal and malignant immortalized human breast cell lines.⁵³ Normal cells, carcinoma cells, mixed normal and carcinoma cells were used. Differences in the spectral backgrounds between the training and test data were observed, which confounds the reproducibility of recorded spectra and, thus, causes the classifier to fail. Using rejection thresholds in the application of the ANN classifier was reported to be helpful in identifying doubtful classifications. Another study⁵⁴ reported imaging fibroadenoma, a benign breast tumor. Data were evaluated using unsupervised cluster analysis by utilizing two spectral regions, namely 1000-1500 and 2800-3000 cm^{-1} . The distribution of four main tissue components-epithelium, retro nuclear basal epithelial regions, mantle zone and distant connective tissue were visualized. The spectral features from each component were discussed in detail. Furthermore, comparing epithelia from fibroadenoma and DCIS, the authors determined that subtle distinctions between the IR characteristics of these two are reproducible. The initial study used tissue from a single patient.

The work was recently extended⁵⁵ to diagnose benign and malignant lesions from 22 patients. The study utilized only spectra from well-defined tumor areas owing to the heterogeneity of tissues. Based on the cluster analysis and on comparison with the H & E images, four classes of distinct breast tissue spectra were identified - fibroadenoma (FA), ductal carcinoma in situ (DCIS), connective tissue and adipose tissue. Further, ANNs were developed as an automated classifier to differentiate the four classes. All spectra of connective tissue and adipose tissue were classified correctly, where the spectral features are clearly different from each other and from tumors as well. Differentiating fibroadenoma from DCIS was more difficult. A toplevel/sublevel strategy was further applied and was able to differentiate 93% between fibroadenoma and DCIS spectra by employing principal component analysis. From the mean spectra, it was found that the DCIS has more lipid content than the fibroadenoma. Invasive ductal carcinoma (IDC) could not be well characterized due to contamination from surrounding cells, illustrating the limited spatial resolution.

Cervical Cancer

The cervix is the lower part of the uterus (womb) in which two major types of cancers occur: squamous cell carcinoma and adenocarcinoma. About 80% to 90% of cervical cancers are squamous cell carcinomas, and the remaining 10% to 20% are adenocarcinomas. Less commonly, cervical cancers have features of both squamous cell carcinomas and adenocarcinomas. These are called adenosquamous carcinomas or mixed carcinomas. Typically, the Papanicolaou (Pap) test checks for changes in the exfoliated cells of cervix to find the presence of any infection, abnormal (unhealthy) cervical cells, or cervical cancer. FTIR spectroscopy, microspectroscopy and FTIR imaging have been widely utilized to study cervical cancer and to perform the same function using computer analyses of spectra.^{26, 56-60} While the first reports in diagnosing cervical cancer are now generally not regarded as leading to solutions,⁵⁶ two groups have provided definitive proof of the potential of IR spectroscopy by careful microscopy studies.^{26, 57, 45, 59, 60} While FTIR images of the amide I and $\nu_{\text{asy}} \text{PO}_2^-$ bands with H&E stained image were compared and only a rough correlation with the pathological features or cell types were obtained, cluster

maps of two, five and eight clusters resulting from UHC analysis for the whole spectrum demonstrated good segmentation. In five clusters, most cell types are apparent including superficial (1), intermediate (2), parabasal (3), and connective tissue (5) upon correlation with the stained image. As in univariate images, the connective tissue region (5) is split into two clusters. Furthermore, by comparing between the UHC analysis of the whole spectrum and only the amide I region, authors demonstrated that minimizing the spectral region for analysis and using fewer clusters does not lead to the loss of useful information. Both univariate FTIR and multivariate images of the sample with several endocervical ducts within the connective tissue were shown. These endocervical ducts lined with columnar endocervical cells were apparent in all those images, in particular even with two clusters.

Cultures derived from cervical cancer cells (HeLa) are one of the most popular model systems and have been studied using FTIR imaging.⁶¹ The cells were directly grown as sparse monolayers onto low-e slides. FTIR image of amide I band region was shown; where large differences in spectral intensities associated with the cells were observed even though these cells are from a homogeneous and exponential cell culture. Cluster analyses of normalized spectra shows distinct differences that were not appreciated in the univariate image. Similarly,⁶² IR imaging with fuzzy C-means clustering and hierarchical cluster analysis were utilized to study the thin sections of cervix uteri encompassing normal, precancerous and squamous cell carcinoma. These studies demonstrate that IR imaging, in combination with multivariate techniques, is capable of segmenting cervical tissues in a manner that is comparable to H&E stained image differentiation and is significantly more sensitive in terms of the chemical composition of the cells – whether it be due to metabolic or disease reasons.

Prostate

Prostate cancer is the most prevalent internal cancer in the US.⁶³ Hence, its pathologic diagnosis and correct interpretation of disease state is crucial.⁶⁴ FTIR imaging has been proposed as a solution that can potentially help pathologists by providing an objective and reproducible assessment of disease in a manner that is easily understood by clinicians. It is also a good model system for the development of FTIR imaging protocols. We first review progress in the field and then describe efforts in our and collaborator's laboratories towards formulating a practical algorithm for prostate cancer pathology. While a number of studies examined human prostate tissue with IR spectroscopy⁶⁵⁻⁶⁸ microscopy approaches have recently been extensively utilized to study both fundamental properties of prostate tissue and to determine structural units in normal and disease states.⁶⁹⁻⁷⁵ An understanding of the tissue is now emerging as a result of these studies. While the fundamental properties of the tissue are being examined, we have focused on developed statistically validated diagnostic methods.

We have utilized high throughput imaging with the express purpose of correlating spectra to clinical practice.^{39, 64, 76} It is instructive to first examine the approaches of some previous studies and then describe our approach in some detail. A variety of techniques have been reported for analyzing prostate tissue, including unsupervised multivariate data analysis techniques such as agglomerative hierarchical clustering (AH), fuzzy C-means (FCM), or k-means (KM) clustering to construct infrared spectral maps of tissue structures.⁷⁷ The results from these multivariate techniques confirmed the standard histopathological techniques and found out to be helpful for identifying and discriminating the tissues structures. Agglomerative hierarchical clustering was

found to be the best method among the cluster imaging methods in terms of segmenting the tissue. While these techniques comprise one end of the approach in using large spectral regions and completely objective methods, the other extreme has also proven to be useful. In the second paradigm, careful examination of the spectral data yields some measures that prove useful. For example, the ratio of peak areas at 1030 and 1080 cm^{-1} , corresponding to the glycogen and phosphate vibrations respectively were utilized as a diagnostic marker for the differentiation of benign from malignant cells.⁶⁹ Authors summarized that the use of this ratio in association with FTIR spectral imaging provides a basis for estimating areas of malignant tissue within defined regions of a specimen. While it may be argued that the former is not based on clinical knowledge and is more suited for discovery, it also involves the choice of selecting specific number of clusters and their subsequent interpretation. The latter is based on a single parameter whose utility for universal diagnoses remains to be tested. Nevertheless, these studies indicate that both approaches provide information about the tissue that is useful.

Our approach has used elements from both pattern recognition and spectroscopic analyses of univariate measures.^{39, 76} In all cases, one starts with the acquired imaging data (figure 4). Since the data set is large (typically 10-1000 GB), it is advisable to reduce the dimensionality of data using some numerical procedure. Compression algorithms, principal components analyses or simply storing only the information needed for classification (if the algorithm is known) is useful. We sought expressly to relate the recorded IR imaging data to clinical knowledge base. Hence we started with a model that is derived from clinical practice. Clearly, the approach limits the discovery of new knowledge but it assures the clinician that all quantities of importance for diagnoses will be considered. The acquired data is labeled with known cell identity or disease states. These pixels are best identified by a combination of very careful manual labeling and test for absorbance fidelity.⁷⁸ Spectra from the label regions are employed via average values, medians and standard deviation analyses to determine a set of spectral features that are descriptive of the major features of all spectra. We first note that the characteristic IR absorbance spectra of ten histological classes comprising prostate tissue look similar. Though small differences in spectral features were observed at many frequencies, summary statistics are limited in their examination of spectra for classification. Further, the small differences indicate that noise and biological variability may render univariate measures less reliable. The large number of classes usually implies that univariate analyses cannot distinguish all histological classes present in the tissues and hence the need for multivariate analyses is apparent. Here the similarity of the spectral features for all classes works in our favor. Very similar baseline points are obtained from an analysis of all spectra and only subtle feature differences are noted to distinguish the various class spectra. Hence, unknown spectra can be processed in the same specified manner, without introducing any bias. Each of these features is termed a metric to denote that it is a useful measure of the spectrum. Individual metrics can allow segmentation of various tissue types if they are sufficiently different in a sampled population.

We then employ the equivalent of a t-test in that the overlap between the absorbance distributions of metrics is determined and equated to the error in prediction. The metrics are arranged in the order of increasing overlap. Hence, we have an ordered set that differentiates at least two classes. To obtain overall accuracy, we employ a modified Bayesian algorithm to provide the probability of each class for every pixel. This fuzzy result is employed to determine the area under the curve (AUC) of a receiver operating characteristic (ROC) curve. The ROC

curve is built from accepting the probability of each class at an increasing threshold that varies between 0 and 1. For optimized threshold values, the fuzzy classification is turned in to a classified image, where each pixel is assigned a distinct class. We note that the method incorporates analysis of all spectral features, a selection of the best features based on statistical analysis of data and an optimal prediction of the class of each pixel based on an objective selection rule from the fuzzy classification. The method is very powerful in that it employs spectral features that are ordinarily employed by spectroscopists as metrics, which permits a spectroscopic analysis of the basis of decision-making. Further, the method explicitly obtains the fuzzy rule data for final classification. The value of the rule data for each class is actually the probability of belonging to the class without consideration for the prior prevalence of the class. Hence, the method can allow direct comparisons between performances for different classes. The dependence of the process on various experimental parameters has also been reported.

The complication inherent in translating the results from small data set of patients to clinical applications is well recognized in the spectroscopy community. The variability in data, arising from variations within and between patients, sample preparation and handling, is likely to provide noisy estimates of performance. Hence, statistical stability may be obtained by examining a large number of samples. Similarly, large number of patients may be employed to provide calibration models, likely improving the robustness of the developed algorithm. We have described a high throughput sampling method from tissues.^{14, 39, 76} Briefly, the approach uses a combinatorial sampling of tissue type and pathology to first acquire small sections of tissues from large archival cases. These small sections are arranged in a grid pattern and placed on the same substrate. The sample is termed a tissue microarray to reflect the similarity with cDNA microarrays. For spectroscopic imaging and the development of automated algorithms, the approach represents a large number of cases that can be used both for accurate prediction algorithm building and for extensive validations. The same approach is likely to prove useful for extensions to determining pathology. Figure 5 demonstrates the typical workflow of a validation algorithm and methods used for statistical comparison. We strongly suggest a variety of methods for measuring performance as each method has its own advantages and disadvantages. For example, summary measures from ROC curves only provide information about accuracy but do not provide which class the inaccuracies arise from. Similarly, confusion matrices provide cross-class information but do not provide global performance measures in the mold of ROC curves.

OUTLOOK

FTIR imaging has experienced rapid growth in the past 10 years and is increasingly being applied to biomedical tissue, especially for the analyses of cancer. The major trends emerging in instrumentation include faster detectors and novel modes of data collection (e.g. time-resolved imaging), of sampling (e.g. ATR) and application areas. For biomedical samples, the information content is quite rich and is often available through simple univariate analysis. For more complex applications, e.g. cancer diagnoses, the data acquisition, sampling and data analyses must be integrated in a coherent manner to provide a practical solution. We anticipate that the technology and its application to biomedical problem will continue to grow with the cooperation of instrument manufacturers, applications scientists, numerical methods developers and communities that can utilize the information effectively, e.g. pathologists or surgeons.

References

1. Gore, R. C.; Barnes, R. B.; Petersen, E., Infrared Absorption of Aqueous Solutions of Organic Acids and Their Salts. *Anal. Chem.* **1949**, 21, (3), 382-386.
2. Kwiatkoski, J. M.; Reffner, J. A., FT-IR microscopy advances. *Nature* **1987**, 328, (6133), 837-838.
3. Koenig, J. L., *Microspectroscopic imaging of polymers*. American Chemical Society Washington, DC: 1998.
4. Williams, D. K.; Schwartz, R. L.; Bartick, E. G., Analysis of latent fingerprint deposits by infrared microspectroscopy. *Appl Spectrosc* **2004**, 58, (3), 313-6.
5. Petrich, W., Mid-Infrared and Raman Spectroscopy for Medical Diagnostics *Applied Spectroscopy Reviews* **2001**, 36, (2), 181-237.
6. Naumann, D., FT-infrared and FT-Raman Spectroscopy in Biomedical Research *Applied Spectroscopy Reviews* **2001**, 36, (2), 239-298.
7. Lewis, E. N.; Levin, I. W., Vibrational Spectroscopic Microscopy: Raman, Near-Infrared and Mid-Infrared Imaging Techniques. *Microscopy and Microanalysis* **1995**, 1, (01), 35-46.
8. Lewis, E. N.; Treado, P. J.; Reeder, R. C.; Story, G. M.; Dowrey, A. E.; Marcott, C.; Levin, I. W., Fourier transform spectroscopic imaging using an infrared focal-plane array detector. *Anal Chem* **1995**, 67, (19), 3377-81.
9. Bhargava, R.; Wang, S. Q.; Koenig, J. L., FTIR microscopy of polymeric systems. *Advances in Polymer Science* **2003**, 163, 137-191.
10. Mendelsohn, R.; Flach, C. R.; Moore, D. J., Determination of molecular conformation and permeation in skin via IR spectroscopy, microscopy, and imaging. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2006**, 1758, (7), 923-933.
11. Garidel, P.; Boese, M., Mid infrared microscopy: mapping and imaging: A bio-analytical tool for spatially and chemically resolved tissue characterization and evaluation of drug permeation within tissues. *Microsc Res Tech* **2007**, 70, (4), 336-349.
12. Lewis, E. N.; Gorbach, A. M.; Marcott, C.; Levin, I. W., High-fidelity Fourier transform infrared spectroscopic imaging of primate brain tissue. *Applied Spectroscopy* **1996**, 50, (2), 263-269.
13. Lewis, E. N.; Kidder, L. H.; Levin, I. W.; Kalasinsky, V. F.; Hanig, J. P.; Lester, D. S., Applications of Fourier transform infrared imaging microscopy in neurotoxicity. *Ann N Y Acad Sci* **1997**, 820, 234-46; discussion 246-7.
14. Levin, I. W.; Bhargava, R., Fourier transform infrared vibrational spectroscopic imaging: integrating microscopy and molecular recognition. *Annu Rev Phys Chem* **2005**, 56, 429-74.
15. Snively, C. M.; Koenig, J. L., Characterizing the Performance of a Fast FT-IR Imaging Spectrometer. *Applied Spectroscopy* **1999**, 53, (2), 170-177.
16. Bhargava, R.; Levin, I. W., Fourier transform infrared imaging: theory and practice. *Anal. Chem* **2001**, 73, (21), 5157-5167.
17. Snively, C. M.; Katzenberger, S.; Oskarsdottir, G.; Lauterbach, J., Fourier-transform infrared imaging using a rapid-scan spectrometer. *Opt. Lett* **1999**, 24, 1841-1843.
18. Huffman, S. W.; Bhargava, R.; Levin, I. W., Generalized Implementation of Rapid-Scan Fourier Transform Infrared Spectroscopic Imaging. *Applied Spectroscopy* **2002**, 56, (8), 965-969.
19. *Spectrochemical Analysis using Infrared Multichannel Detectors*, Bhargava, R.

Levin, I. W. (Eds). Blackwell Publishing, Sheffield, UK: 2005.

20. Bhargava, R.; Levin, I. W., Time-resolved Fourier transform infrared spectroscopic imaging. *Appl Spectrosc* **2003**, 57, (4), 357-66.

21. Bhargava, R.; Levin, I. W., Noninvasive Imaging of Molecular Dynamics in Heterogeneous Materials. *Macromolecules* **2003**, 36, (1), 92-96.

22. Bhargava, R.; Levin, I. W., Gram-Schmidt orthogonalization for rapid reconstructions of Fourier transform infrared spectroscopic imaging data. *Appl Spectrosc* **2004**, 58, (8), 995-1000.

23. Bhargava, R.; Levin, I. W., Enhanced Time-Resolved Fourier Transform Infrared Spectroscopic Imaging for Reversible Dynamics. *J. Phys. Chem. A* **2004**, 108, (18), 3896-3901.

24. O'Leary, T. J.; Engler, W. F.; Ventre, K. M., Infrared Microspectroscopy of Human Tissue. *Applied Spectroscopy* **1989**, 43, (6), 1095-1097.

25. Marcott, C.; Story, G. M.; Dukor, R. K., Infrared Spectral Imaging of H&E-Stained Breast Tissue Biopsies. *Microscopy and Microanalysis* **2004**, 10, (S02), 182-183.

26. Romeo, M.; Mohlenhoff, B.; Jennings, M.; Diem, M., Infrared micro-spectroscopic studies of epithelial cells. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2006**, 1758, (7), 915-922.

27. Patterson, B. M.; Havrilla, G. J., Attenuated Total Internal Reflection Infrared Microspectroscopic Imaging Using a Large-Radius Germanium Internal Reflection Element and a Linear Array Detector. *Applied Spectroscopy* **2006**, 60, (11), 1256-1266.

28. Sommer, A. J.; Tisinger, L. G.; Marcott, C.; Story, G. M., Attenuated Total Internal Reflection Infrared Mapping Microspectroscopy Using an Imaging Microscope. *Applied Spectroscopy* **2001**, 55, (3), 252-256.

29. Otts, D. B.; Zhang, P.; Urban, M. W., High Fidelity Surface Chemical Imaging at 1000 nm Levels: Internal Reflection IR Imaging (IRIRI) Approach. *Langmuir* **2002**, 18, (17), 6473-6477.

30. Chan, K. L. A.; Kazarian, S. G., New opportunities in micro- and macro-attenuated total reflection infrared spectroscopic imaging: Spatial resolution and sampling versatility. *Applied Spectroscopy* **2003**, 57, (4), 381-389.

31. Chan, K. L.; Hammond, S. V.; Kazarian, S. G., Applications of attenuated total reflection infrared spectroscopic imaging to pharmaceutical formulations. *Anal Chem* **2003**, 75, (9), 2140-6.

32. Colley, C. S.; Kazarian, S. G.; Weinberg, P. D.; Lever, M. J., Spectroscopic imaging of arteries and atherosclerotic plaques. *Biopolymers* **2004**, 74, (4), 328-35.

33. Kazarian, S. G.; Chan, K. L. A., Applications of ATR-FTIR spectroscopic imaging to biomedical samples. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2006**, 1758, (7), 858-867.

34. van der Weerd, J.; Andrew Chan, K. L.; Kazarian, S. G., An innovative design of compaction cell for in situ FT-IR imaging of tablet dissolution. *Vibrational Spectroscopy* **2004**, 35, (1-2), 9-13.

35. van der Weerd, J.; Kazarian, S. G., Combined approach of FTIR imaging and conventional dissolution tests applied to drug release. *J Control Release* **2004**, 98, (2), 295-305.

36. Chan, K. L.; Kazarian, S. G., Fourier transform infrared imaging for high-throughput analysis of pharmaceutical formulations. *J Comb Chem* **2005**, 7, (2), 185-9.

37. Chan, K. L.; Kazarian, S. G., ATR-FTIR spectroscopic imaging with expanded field of view to study formulations and dissolution. *Lab Chip* **2006**, 6, (7), 864-70.

38. Snively, C. M.; Oskarsdottir, G.; Lauterbach, J., Chemically sensitive parallel analysis of combinatorial catalyst libraries. *Catalysis Today* **2001**, 67, (4), 357-368.

39. Fernandez, D. C.; Bhargava, R.; Hewitt, S. M. ; Levin, I. W. , Infrared spectroscopic imaging for histopathologic recognition. *Nat Biotechnol* **2005**, 23, (4), 469-74.
40. Boskey, A. L.; Mendelsohn, R., Infrared spectroscopic characterization of mineralized tissues. *Vibrational Spectroscopy* **2005**, 38, (1-2), 107-114.
41. Posner, A. S.; Duyckaerts, G., Infrared study of the carbonate in bone, teeth and francolite. *Experientia* **1954**, 10, (10), 424-5.
42. Boskey, A.; Mendelsohn, R., Infrared analysis of bone in health and disease. *J Biomed Opt* **2005**, 10, (3), 031102.
43. Boskey, A. L.; Goldberg, M.; Kulkarni, A.; Gomez, S., Infrared imaging microscopy of bone: Illustrations from a mouse model of Fabry disease. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2006**, 1758, (7), 942-947.
44. Faibish, D.; Gomes, A.; Boivin, G.; Binderman, I.; Boskey, A., Infrared imaging of calcified tissue in bone biopsies from adults with osteomalacia. *Bone* **2005**, 36, (1), 6-12.
45. Bamberg, K. R.; Schultke, E.; Wood, B. R.; MacDonald, S. T. R.; Ataelmannan, K.; Griebel, R. W. ; Juurlink, B. H. J.; McNaughton, D., A Fourier transform infrared microspectroscopic imaging investigation into an animal model exhibiting glioblastoma multiforme. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2006**, 1758, (7), 900-907.
46. American Brain Tumor Association.
47. Krafft, C.; Shapoval, L.; Sobottka, S. B.; Schackert, G.; Salzer, R., Identification of primary tumors of brain metastases by infrared spectroscopic imaging and linear discriminant analysis. *Technol Cancer Res Treat* **2006**, 5, (3), 291-8.
48. Amharref, N.; Beljebbar, A.; Dukic, S.; Venturolo, L.; Schneider, L.; Pluot, M.; Vistelle, R.; Manfait, M., Brain tissue characterisation by infrared imaging in a rat glioma model. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2006**, 1758, (7), 892-899.
49. Kneipp, J.; Lasch, P.; Baldauf, E.; Beekes, M.; Naumann, D., Detection of pathological molecular alterations in scrapie-infected hamster brain by Fourier transform infrared (FT-IR) spectroscopy. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **2000**, 1501, (2-3), 189-199.
50. Kidder, L. H.; Kalasinsky, V. F.; Luke, J. L.; Levin, I. W.; Lewis, E. N., Visualization of silicone gel in human breast tissue using new infrared imaging spectroscopy. *Nat Med* **1997**, 3, (2), 235-237.
51. Fabian, H.; Lasch, P.; Boese, M.; Haensch, W., Mid-IR microspectroscopic imaging of breast tumor tissue sections. *Biopolymers - Biospectroscopy Section* **2002**, 67, (4-5), 354-357.
52. Bhargava, R.; Wall, B. G.; Koenig, J. L. , Comparison of the FT-IR mapping and imaging techniques applied to polymeric systems. *Applied Spectroscopy* **2000**, 54, 470-479.
53. Zhang, L.; Small, G. W.; Haka, A. S.; Kidder, L. H.; Lewis, E. N., Classification of Fourier transform infrared microscopic imaging data of human breast cells by cluster analysis and artificial neural networks. *Appl Spectrosc* **2003**, 57, (1), 14-22.
54. Fabian, H.; Lasch, P.; Boese, M.; Haensch, W. , Infrared microspectroscopic imaging of benign breast tumor tissue sections. *Journal of Molecular Structure* **2003**, 661-662, (1-3), 411-417.
55. Fabian, H.; Thi, N. A. N.; Eiden, M.; Lasch, P.; Schmitt, J.; Naumann, D., Diagnosing benign and malignant lesions in breast tissue sections by using IR-microspectroscopy. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2006**, 1758, (7), 874-882.

56. Wong, P. T.; Wong, R. K.; Caputo, T. A.; Godwin, T. A.; Rigas, B., Infrared spectroscopy of exfoliated human cervical cells: evidence of extensive structural changes during carcinogenesis. *Proc Natl Acad Sci U S A* **1991**, 88, (24), 10988-92.
57. Boydston-White, S.; Romeo, M.; Chernenko, T.; Regina, A.; Miljkovic, M.; Diem, M., Cell-cycle-dependent variations in FTIR micro-spectra of single proliferating HeLa cells: Principal component and artificial neural network analysis. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2006**, 1758, (7), 908-914.
58. Walsh, M. J.; Germann, M. J.; Singh, M.; Pollock, H. M.; Hammiche, A.; Kyrgiou, M.; Stringfellow, H. F.; Paraskevaidis, E.; Martin-Hirsch, P. L.; Martin, F. L., IR microspectroscopy: potential applications in cervical cancer screening. *Cancer Letters* **2007**, 246, (1-2), 1-11.
59. Bambery, K. R.; Wood, B. R.; Quinn, M. A.; McNaughton, D., Fourier transform infrared imaging and unsupervised hierarchical clustering applied to cervical biopsies. *Australian Journal of Chemistry* **2004**, 57, (12), 1139-1143.
60. Wood, B. R.; Chiriboga, L.; Yee, H.; Quinn, M. A.; McNaughton, D.; Diem, M., Fourier transform infrared (FTIR) spectral mapping of the cervical transformation zone, and dysplastic squamous epithelium. *Gynecol Oncol* **2004**, 93, (1), 59-68.
61. Diem, M.; Romeo, M. J.; Boydston-White, S.; Matthaus, C., "IR Spectroscopic Imaging: from Cells to Tissue" in "Spectrochemical Analysis using Infrared Multichannel Detectors", R. Bhargava and I. W. Levin, Editors, Blackwell Publishing, Sheffield, UK: 2005.
62. Steller, W.; Eienkel, J.; Horn, L. C.; Braumann, U. D.; Binder, H.; Salzer, R.; Krafft, C., Delimitation of squamous cell cervical carcinoma using infrared microspectroscopic imaging. *Anal Bioanal Chem* **2006**, 384, (1), 145-54.
63. http://seer.cancer.gov/csr/1975_2004/results_single/sect_01_table.01.pdf. **2007**.
64. Bhargava, R., *Anal Bioanal Chem (in press)*.
65. Paluszkiwicz, C.; Kwiatek, W. M., Analysis of human cancer prostate tissues using FTIR microspectroscopy and SRIXE techniques. *Journal of Molecular Structure* **2001**, 565, 329-334.
66. Bhargava, R.; Fernandez, D. C.; Schaeberle, M. D.; Levin, I. W., Theory and application of gain ranging to Fourier transform infrared spectroscopic Imaging. *Applied Spectroscopy* **2001**, 55, (12), 1580-1589.
67. Hsu, H. S.; Lin, S. Y.; Li, M. J.; Liang, R. C., Ultrastructural and biophysical studies on protein conformations of epithelium and stroma in benign prostatic hyperplasia before and after transurethral resection of the prostate. *Ultrastructural Pathology* **2002**, 26, (3), 137-141.
68. Li, M. J.; Hsu, H. S.; Liang, R. C.; Lin, S. Y., Infrared microspectroscopic detection of epithelial and stromal growth in the human benign prostatic hyperplasia. *Ultrastructural Pathology* **2002**, 26, (6), 365-370.
69. Gazi, E.; Dwyer, J.; Gardner, P.; Ghanbari-Siahkali, A.; Wade, A. P.; Miyan, J.; Lockyer, N. P.; Vickerman, J. C.; Clarke, N. W.; Shanks, J. H.; Scott, L. J.; Hart, C. A.; Brown, M., Applications of Fourier transform infrared microspectroscopy in studies of benign prostate and prostate cancer. A pilot study. *J Pathol* **2003**, 201, (1), 99-108.
70. Gazi, E.; Dwyer, J.; Lockyer, N.; Gardner, P.; Vickerman, J. C.; Miyan, J.; Hart, C. A.; Brown, M.; Shanks, J. H.; Clarke, N., The combined application of FTIR microspectroscopy and ToF-SIMS imaging in the study of prostate cancer. *Faraday Discuss* **2004**, 126, 41-59; discussion 77-92.
71. Gazi, E.; Dwyer, J.; Lockyer, N. P.; Miyan, J.; Gardner, P.; Hart, C.; Brown, M.; Clarke, N. W., Fixation protocols for subcellular imaging by synchrotron-based Fourier transform infrared microspectroscopy. *Biopolymers* **2005**, 77, (1), 18-30.

72. Gazi, E.; Dwyer, J.; Lockyer, N.; Gardner, P.; Miyan, J.; Hart, C.; Brown, M.; Clarke, N., A study of cytokeratin and mucin in prostate cancer cells using synchrotron based FTIR-microspectroscopic imaging. *Vibrational spectroscopy* **2005**, 38, (1-2), 193-201.
73. German, M. J.; Hammiche, A.; Ragavan, N.; Tobin, M. J.; Cooper, L. J.; Matanhelia, S. S.; Hindley, A. C.; Nicholson, C. M.; Fullwood, N. J.; Pollock, H. M.; Martin, F. L., Infrared spectroscopy with multivariate analysis potentially facilitates the segregation of different types of prostate cell. *Biophysical Journal* **2006**, 90, (10), 3783-3795.
74. Gazi, E.; Baker, M.; Dwyer, J.; Lockyer, N. P.; Gardner, P.; Shanks, J. H.; Reeve, R. S.; Hart, C. A.; Clarke, N. W.; Brown, M. D., A correlation of FTIR spectra derived from prostate cancer biopsies with Gleason grade and tumour stage. *European Urology* **2006**, 50, (4), 750-761.
75. Wolkers, W. F.; Balasubramanian, S. K.; Ongstad, E. L.; Zec, H. C.; Bischof, J. C., Effects of freezing on membranes and proteins in LNCaP prostate tumor cells. *Biochimica Et Biophysica Acta-Biomembranes* **2007**, 1768, (3), 728-736.
76. Bhargava, R.; Fernandez, D. C.; Hewitt, S. M.; Levin, I. W., High throughput assessment of cells and tissues : Bayesian classification of spectral metrics from infrared vibrational spectroscopic imaging data. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **2006**, 1758, (7), 830-845.
77. Lasch, P.; Diem, M.; Naumann, D. In *FT-IR microspectroscopic imaging of prostate tissue sections*, Biomedical Vibrational Spectroscopy and Biohazard Detection Technologies, San Jose, CA, USA, 2004; SPIE: San Jose, CA, USA, 2004; pp 1-9.
78. Bhargava, R.; Hewitt, S. M.; Levin, I. W., Reply to Unrealistic expectations for IR microspectroscopic imaging. *Nat Biotech* **2007**, 25, (1), 31-33.

SBFP Javelin (Rolling Mode)

- 64 x 64 bump-bonded: 180 Hz (1996)
- 64 x 64 : 250 Hz to 315 Hz (1996-98)
- 64 x 64 : 430 Hz (2000)
- 64 x 64 : 615 Hz, Triggered Mode (2001)
- Step-scan (1997); rapid-scan (1999) imaging

Perkin-Elmer Spotlight

- 16 x 1 “linear” array (2001)
- ~ 100 spectra/s
- Exceptionally high SNR
- Thermo: 16 x 2 array (2005)

NIH256-SBFP (Snapshot Mode)

- 256 x 256 MBE grown (1997) - NIST
- 256 x 256 : 143 Hz Capable
- Rapid-scan (2000) imaging
- TRS : 10 ms (2002)
- TRS : 0.1 ms resolution (2004)

FBI-NIH128-RSC (Snapshot)

- 128 x 128 MBE: >16 KHz (2003)
- On-chip co-addition
- Advanced software
- Spatial Subset
- Trigger
- Rapid Scan Imaging (2005)
- Potential
 - Rapid scan : 0.06 s acquisition
 - TRS : 40 microsecond
 - Step-scan : High SNR

Digilab-SBFP Lancer (Snapshot)

- 64 x 64 MBE: 3774 Hz (2002)
- Step-scan imaging (2002)
- Digilab “fast” scan ~ 10 s acquisition
- NIH/Akron rapid scan : 0.25 s acquisition

Figure 1. Various MCT FPA detectors employed for FTIR imaging since the first reports using Santa Barbara Focalplane (SBFP) array detectors. The years in parentheses are the first reports of use for FTIR imaging. Perkin-Elmer introduced the concept of utilizing a small linear array for very high signal to noise ratios, an approach that has since been adopted by Thermo. Our research efforts have involved the use of a high end, custom-built detector that allows for fast imaging.

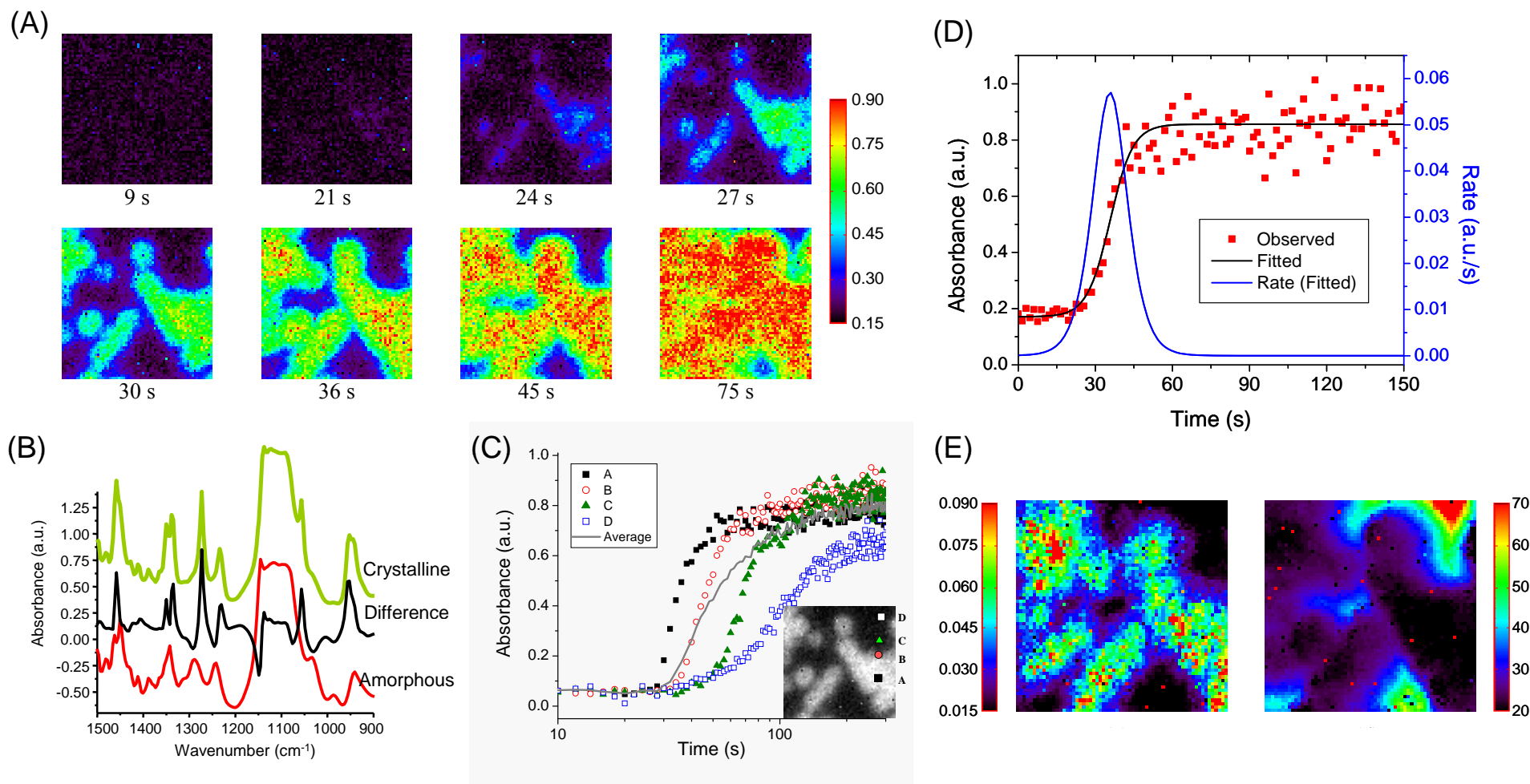


Figure 2. FTIR spectroscopy and imaging permits examination of molecular conformation changes preceding and during polymer crystallization. (A) The distribution of crystalline and amorphous fractions as a function of time for undercooling PEO $\sim 13^\circ\text{C}$ below its melting point can be observed by the intensity of any peak that is different (B). The pixels crystallizing first can be analyzed prior to crystal formation for pre-ordering transitions. (C) Different regions of the sample have different kinetics (symbols), which are not apparent in the average spectral change (line). (D) The kinetic data (noisy) can be fit with a smooth curve and the rate of crystallization obtained. (E) spatial variation of crystallization rate (E, left) correlates with the onset of crystallization (E, right). Those regions that start to crystallize late also have a lower rate and lower ultimate purity, likely due to diffusion of impurities.

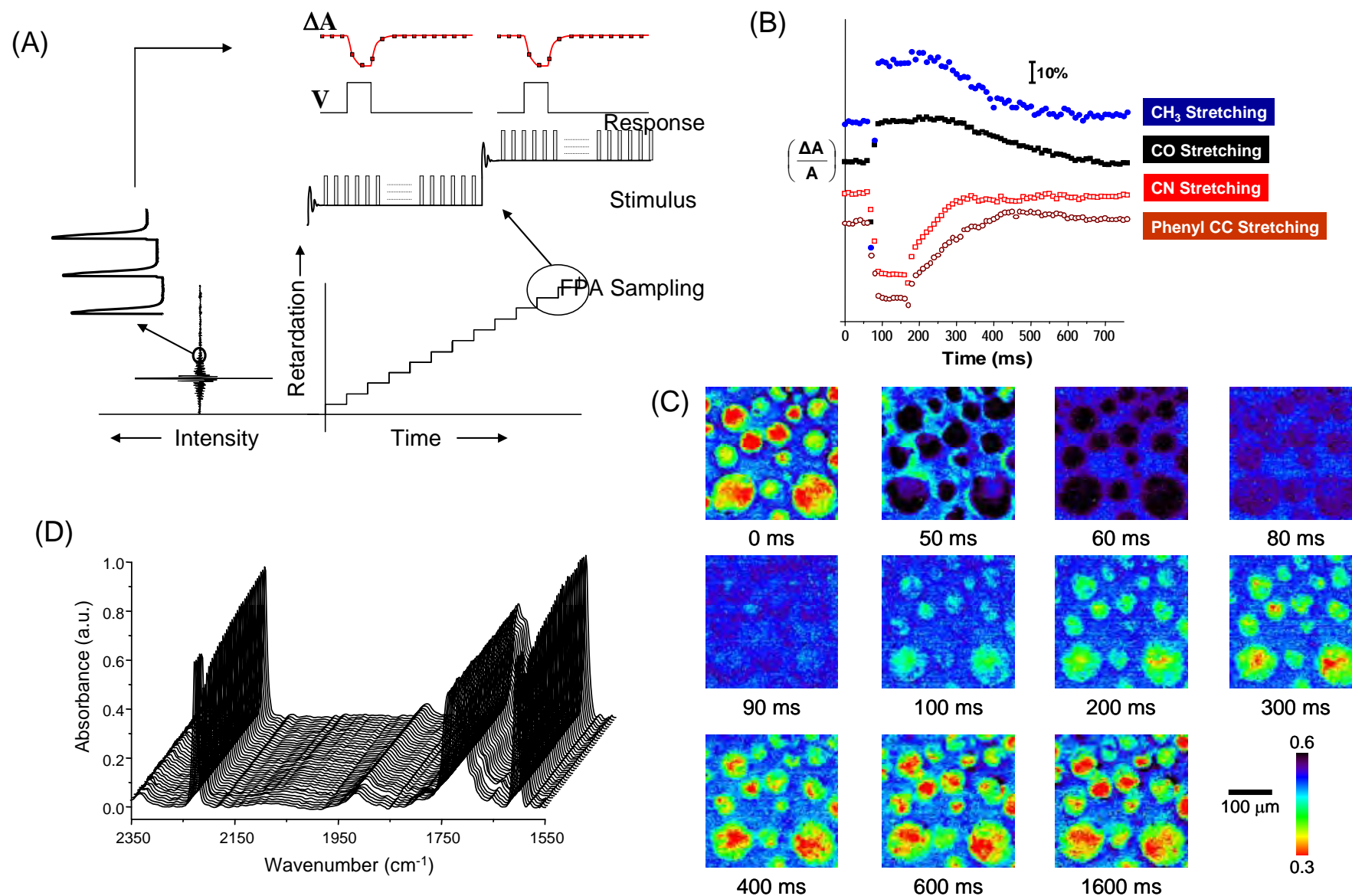


Figure 3. Time-resolved FTIR imaging can provide spatially-resolved, millisecond level dynamics over large sample areas. The operation (A) of the interferometer is similar to that of conventional step scan spectroscopy, except that an entire image is acquired for every sampling point. (B) Various functional groups may be monitored in time at specific pixels or (C) the entire image may be visualized. (D) Entire spectra from pixels may also be observed in the manner of conventional time resolved spectroscopy.

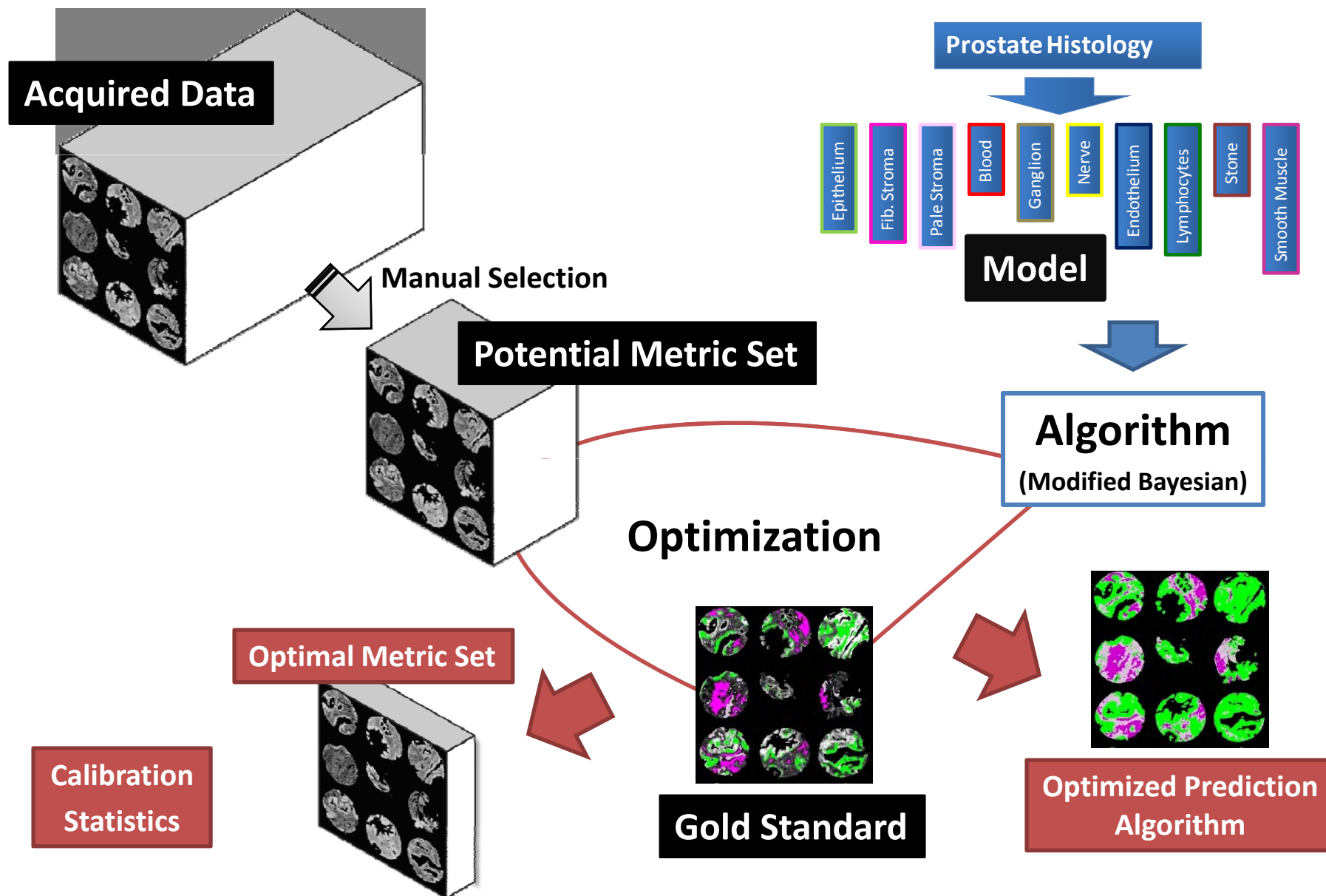
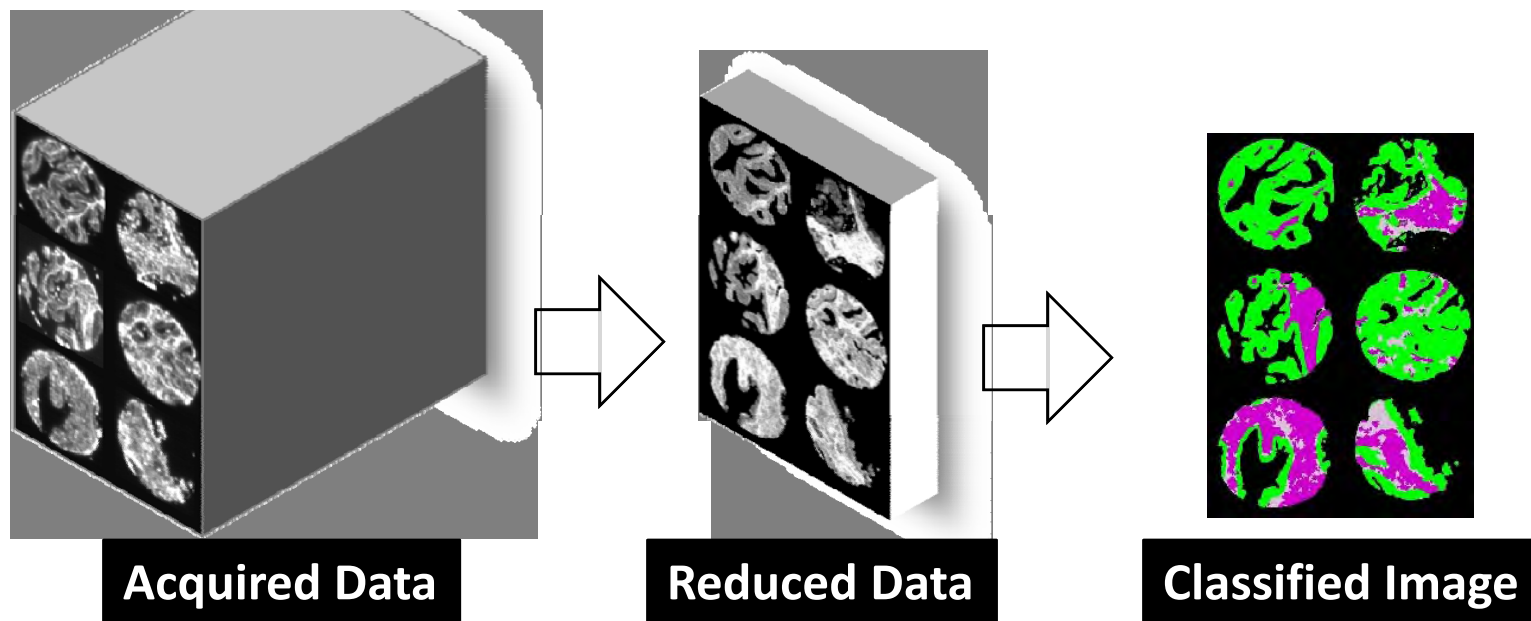
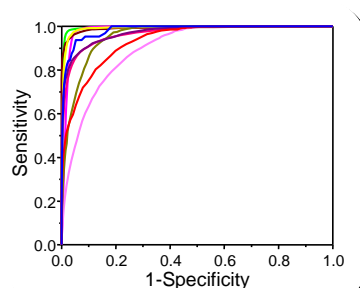


Figure 4. Organization of data into a prediction algorithm involves several steps. Acquired FTIR imaging data (top, left) is reduced by manual selection to a set of features that capture the essential elements of spectra from all tissue types. A model (top, right) is selected for the data and employed to develop an algorithm. The algorithm is applied to the entire metric set and prediction capabilities are optimized. Results of the optimization provide an optimal metric set for validation studies, the parameters of the algorithm to be applied and calibration classification statistics. The optimized algorithm is applied to acquired data without supervision (figure 5).



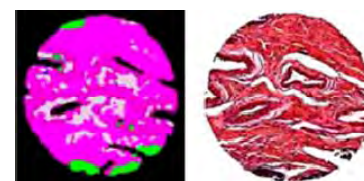
Ground Truth Class \ Result of Classification	EPITHELIUM	MIXED STROMA	FIBROUS STROMA	SMOOTH MUSCLE
EPITHELIUM	95.55	0.16	0.21	0.00
MIXED STROMA	0.00	92.51	0.79	2.76
FIBROUS STROMA	0.19	1.15	93.04	0.69
SMOOTH MUSCLE	0.00	5.53	0.94	94.04

Confusion Matrices



ROC Curves

Validation/Statistics



Classified Images

Figure 5. Validation or unsupervised application of the developed protocol. FTIR imaging data are acquired (top, left), reduced to the optimal metric set (obtained as in figure 4), which is then converted to a single image that denotes each cell type by a specific color and empty space by black (top, right). Classified images can be compared to ground truth images by using confusion matrices, ROC curves and by comparisons of pixels between images. Statistical measures from these validation tests provide quantifiable results and high confidence in the development of robust algorithms.

Practical protocols for fast histopathology by Fourier transform infrared spectroscopic imaging

Frances N. Keith, Rohith K. Reddy, and Rohit Bhargava*

Department of Bioengineering and Beckman Institute for Advanced Science and Technology,
University of Illinois at Urbana-Champaign
3216 Digital Computer Lab, MC-278, 1304 W. Springfield Ave., Urbana, IL 61801

ABSTRACT

Fourier transform infrared (FT-IR) spectroscopic imaging is an emerging technique that combines the molecular selectivity of spectroscopy with the spatial specificity of optical microscopy. We demonstrate a new concept in obtaining high fidelity data using commercial array detectors coupled to a microscope and Michelson interferometer. Next, we apply the developed technique to rapidly provide automated histopathologic information for breast cancer. Traditionally, disease diagnoses are based on optical examinations of stained tissue and involve a skilled recognition of morphological patterns of specific cell types (histopathology). Consequently, histopathologic determinations are a time consuming, subjective process with innate intra- and inter-operator variability. Utilizing endogenous molecular contrast inherent in vibrational spectra, specially designed tissue microarrays and pattern recognition of specific biochemical features, we report an integrated algorithm for automated classifications. The developed protocol is objective, statistically significant and, being compatible with current tissue processing procedures, holds potential for routine clinical diagnoses. We first demonstrate that the classification of tissue type (histology) can be accomplished in a manner that is robust and rigorous. Since data quality and classifier performance are linked, we quantify the relationship through our analysis model. Last, we demonstrate the application of the minimum noise fraction (MNF) transform to improve tissue segmentation.

Keywords: Breast Cancer, FT-IR Spectroscopy, Hyperspectral, Histopathology, Imaging, Diagnostics, MNF Transform

1. INTRODUCTION

As histologic analysis of biopsied tissue forms the standard in definitive diagnosis of breast lesions, it is estimated that more than 1.6 million women undergo breast biopsies each year in the US alone. Biopsy samples are fixed to ensure tissue stability¹ and then sectioned for staining.² Microscopic examinations of stained tissue sections by a trained pathologist are the gold standard used in diagnosing breast cancer.³ Unfortunately, these evaluations are time consuming⁴ and do not always lead to an unequivocal diagnosis. For example, a study of 481 breast cancer patients from 1982-2000 at a regional cancer center indicated that 73% of ductal carcinoma *in situ* (DCIS) patients are referred by a general pathologist to an expert pathologist for review.⁵ After review, 43% of these cases received different treatment recommendations. Another study found that 52% of cases referred to a multidisciplinary tumor review board received different surgery recommendations.⁶ Clearly, the diagnostic process is sub-optimal. Rapid, objective second opinions are desirable. The use of emerging biological understanding and technologies for diagnoses could provide additional information in tumor evaluation and help make accurate therapy decisions. Further, it is likely that the morphologic parameters of current diagnoses are insufficient and additional information must be added. This information is typically biochemical in nature. For example, staining for human epidermal growth factor receptor 2 (HER2) can identify 25-30% of breast cancers.⁷ Such examples of success, unfortunately, are uncommon for cancers in complex tissues. Hence, alternative methods are urgently required to aid diagnostic pathology.

One such means is the use of molecular spectroscopy. For example, Fourier transform infrared (FT-IR) spectroscopy is traditionally used for molecular identifications and biomolecular structure elucidations, but is not currently applied in clinical pathology.⁸ An IR spectrum provides a unique molecular fingerprint with a quantitative measure of the molecular bonds present in an examined material.⁹ Thus it should give a reproducible measurement of tissue

* rxb@uiuc.edu; phone 1 217 265 6596; cisl.bioen.uiuc.edu

composition. Tissue, however, is microscopically heterogeneous and the measurement of chemical composition must be made in the context of knowledge of tissue structure (histology).¹⁰ The recent emergence of FT-IR imaging couples spectroscopy and microscopy to permit rapid acquisition of spectra from tens of thousands of pixels at a high spatial resolution. Each pixel (spectrum) typically contains thousands of data points in the mid-IR wavelength region (2-12 μ m).¹¹ Automated classification can then be employed for rapid computerized tissue image analysis, as has been practiced in both the spectral processing and image processing communities. The end goal of the measurement and associated data processing steps is to permit the rapid segmentation of different types of tissue without the need for chemical dyes or contrast agents.¹⁰ Last, the use of FT-IR imaging only involves light interacting with a sample and, unlike conventional biochemical analysis methods, does not alter the tissue in any manner. Thus it can provide additional information for pathology without the necessity of additional materials, tissue samples or changes in clinical protocols.

In this manuscript we use breast tissue as an example to illustrate the application of FT-IR imaging coupled with computerized classification for histopathology. Specifically, we demonstrate that a combination of FT-IR imaging, classification algorithms and integrated computational methods for enhancement of acquired data can be used in tandem to optimize the development of practical protocols for automated histopathology. Previous studies report on the potential for IR spectroscopy in breast pathology,^{12,13,14,15,16,17} but no complete study on the spectral features of different histologic types of breast tissue exists. Preliminary efforts indicate significant spectral variation between different types of breast tissue and breast tumors,^{18,19,20} but a protocol for clinical translation is lacking. We combine fast FT-IR imaging and tissue microarray sampling to demonstrate the effectiveness of our approach for automated breast histopathology on normal and malignant tissue from five patients. This approach is distinct from that in Raman spectroscopy, where histologic models are used in analyzing spectra.^{21,22} As a first step towards automated tissue segmentation, we distinguish breast stroma and epithelium. This is a critical step, as over 99% of breast tumors arise in the epithelial tissue lining milk ducts and lobules.²³ False color classified images denoting stroma and epithelium are produced, followed by analysis of data collection parameters. We evaluate the impact of spectral resolution and noise on classification accuracy to demonstrate potential for faster data acquisition without loss in classification confidence. This study presents an initial effort in developing applications for FT-IR imaging in clinical pathology.

2. METHODOLOGY

2.1 Data Acquisition

The first studies to examine IR spectra of tissue began over fifty years ago,²⁴ but the field did not truly make progress due to limitations in instrumentation. Today, a combination of an IR microscope, Michelson interferometer and focal plane array (FPA) detector²⁵ permits efficient data acquisition for large sample areas. The data presented in this study is collected using the Perkin-Elmer Spotlight 400 imaging spectrometer. A spatial pixel size of 6.25 μ m and a spectral resolution of 4 cm^{-1} were employed, with 2 scans averaged for each pixel. An IR background is collected with 120 scans co-added at a location on the substrate where no tissue is present. No undersampling was employed in data acquisition and a NB medium apodization function was used. A ratio of the background to tissue spectra is then computed to remove substrate and air contributions to the spectral data. The Spotlight software atmospheric correction algorithm is applied to eliminate remaining atmospheric contributions to the tissue spectra. As opposed to other configurations that employ a large FPA detector, this instrument employs a linear array detector that is raster scanned to acquire data from large sample areas. We use a combination of instrument control and post-processing software to computationally re-organize data acquired into large image sizes. Images of stained tissue are acquired using a standard Zeiss optical microscope.

2.2 Tissue sampling

Tissue microarrays (TMAs) permit facile comparison of small tissue samples from numerous patients²⁶ and are an especially useful sampling medium for spectroscopic analyses.²⁷ A TMA contains numerous small round tissue samples, termed cores, which are extracted from biopsy samples from different patients. Two paraffin-embedded TMAs were obtained from a commercial source (US Biomax) for this study. The first TMA section is placed on a glass slide and stained with hematoxylin and eosin (H&E) dyes. In H&E staining, hematoxylin stains nucleic acids and eosin stains protein-rich tissue regions. This section is used for visual morphology interpretation by a pathologist. The second TMA section is placed on a barium fluoride (BaF_2) substrate for FT-IR imaging. Though the arrays contained a large number

of samples, a smaller subset of malignant and normal tissue cores from five patients with invasive ductal carcinoma (IDC) is selected for this study as the illustrative example. Each of the ten cores is 1.5 mm in diameter; hence, at a 6.25 μm pixel size, approximately 280,000 spectra are collected for each core. This results in the collection of over 560,000 spectra for each patient and approximately 2.8 million total spectra for all ten cores. This large spectral dataset facilitates rigorous validation of classification protocols at a pixel level. Paraffin is removed from the TMA by immersion in hexane with continuous stirring at 40 $^{\circ}\text{C}$ for 48-72 hours. Spectra are recorded at several locations on the TMA every 24 hours during this period to monitor paraffin removal with the disappearance of the 1462 cm^{-1} peak.

2.3 Image analysis and classification

A supervised segmentation method is used for FT-IR image classification. This algorithm has been described in detail elsewhere,²⁸ but is based on a modified version of a Bayesian classifier. First, the spectral profile of 1641 bands is reduced to a set of 89 useful metrics by examination of spectra from manually selected stroma and epithelium tissue regions. Metrics are manually selected to include peak ratios, peak areas, and peak centers of gravity. A metric profile M is generated for each pixel in each tissue image of the form

$$M = [m_1, m_2, m_3, \dots, m_{n_m}], n_m = 89 \quad (1)$$

where each m_i is the value for a single metric and n_m is the total number of manually selected metrics. Frequency distributions for stroma and epithelium are determined for each metric and used to estimate the probability of a given metric profile representing either of these two classes. The probability of an image pixel from each class c_i being represented by a given metric profile is determined using Bayes' Rule

$$p(c_i|M) = \frac{p(M|c_i)p(c_i)}{p(M)} \quad (2)$$

where $p(M|c_i)$ is estimated from the metric class frequency distributions and $p(M)$ is the probability of a given metric profile. The prior probability of particular tissue class $p(c_i)$ in this model cannot be determined due the manual selection of tissue classes on FT-IR images, and is estimated as 0.5. Other ways to estimate or optimize the class prior probability may be utilized; we have noticed anecdotally, however, that the choice of this value across a large range does not significantly affect the classification results. Classification accuracy is estimated with receiver operating characteristic (ROC) analysis for selected tissue regions. The area under the ROC curve (AUC) is used to evaluate classifier sensitivity and specificity and estimate the potential of the algorithm for accurate histology determinations. The classification algorithm is trained on a large array dataset and separately validated on a second array. It is notable that we do not develop the entire classification algorithm anew here. First, the central idea of this manuscript is to demonstrate the optimization of a developed protocol and second, the sample sizes chosen here are insufficient for de novo algorithm development. Data is analyzed using the Environment for Visualizing Images (ENVI) software and with programs written in-house using Interactive Data Language (IDL).

2.4 Spectral resolution and noise analysis

Spectral resolution and noise are two common experimental variables that affect results in IR spectral analyses. The effects of spectral resolution and spectral noise are evaluated here in the context of quantitative histologic segmentation to minimize data collection time. As per the trading rules of IR spectroscopy, data collection time is expected to decrease linearly with spectral resolution and a quadratic rate with reduction in signal-to-noise ratio (SNR).²⁹ Ideally, these parameters would be analyzed by acquiring data at different spectral resolutions and numbers of spectral co-adds. However, the time required to collect multiple images for the TMA is prohibitive. Instead, computational methods are used to examine these parameters using the original FT-IR images acquired at 4 cm^{-1} and 2 scans per pixel. First, spectral resolution is evaluated by downsampling the data using a neighbor binning procedure to resolutions of 8, 16, 32, 64 and 128 cm^{-1} . Classification is then performed on downsampled datasets to determine the coarsest spectral resolution needed for satisfactory stroma and epithelium segmentation. For a fine spectral resolution data set at 4 cm^{-1} , the effect of noise is evaluated by adding to each spectrum noise in Gaussian distributions with standard deviations of 0.001, 0.01, and 0.1 au. Classification accuracy is estimated by evaluating the AUC at each noise standard deviation. Computational noise reduction with the minimum noise fraction (MNF) transform³⁰ is evaluated by reducing noise in all the data sets. Classification is performed with the same algorithm on these MNF transformed images to determine the impact of this noise reduction algorithm on stroma and epithelium segmentation.

3. DATA

The classification model presented in this manuscript involves segmentation of stroma and epithelium, which are the two most prominent tissue classes in fixed breast tissue used for pathology evaluation.³¹ In practice, the recognition of epithelial cells is especially critical for cancer diagnoses, as the vast majority (>99%) of breast cancers arise in this cell type.²³ Hence, the two class model is of practical significance. While seemingly simple and practical, however, the model can potentially be confounding as stroma consists of many cell types with disparate spectral characteristics. This model was employed to develop a classifier using training data from a TMA with forty patients. Final model calibration for sixty eight tissue cores yielded an AUC value of 0.99 with an eight metric classifier.^{32,33} In this study we validate this classifier with one malignant and a matched normal TMA core from a subset of five patients. As seen in Figure 1A and B, absorbance images based on spectral features closely compare with images of H&E stained tissue. Hence, using conventional pathology knowledge we can select image pixels that unequivocally correspond between the two images - representing both stroma and epithelium. These pixels are selected by examining FT-IR images at 1080 cm^{-1} to highlight asymmetric PO_2 stretching vibrations in glycoprotein in epithelium,¹⁴ 1236 cm^{-1} to highlight CH_2 wagging vibrations associated with collagen proteins,³⁴ 1652 cm^{-1} to highlight C=O stretching vibrations at the protein amide I mode,³⁴ and 3292 cm^{-1} to highlight NH bending vibrations at the protein amide A mode (shown as an example in Figure 1B).³⁵ We emphasize that multiple vibrational modes must be examined in tandem and pixels identified with great care and diligence as these form the gold standard for future comparisons. Over 185,000 pixels are marked in these ten tissue cores to serve as the gold standard for ROC analysis (as shown in Figure 1C). Selecting this large set of pixels is important to achieve a reasonable sample size to accurately estimate classification potential for the entire data set. Boundary pixels are not marked to avoid errors associated with mixed pixels in FT-IR images.²⁷ A qualitative comparison of stained and classified images indicates that stroma and epithelium segmentation is reasonable (Figure 1D), and this is confirmed with an AUC value of 0.98 after quantitative ROC analysis. Stroma and epithelium are easily identified on false color classified images without detailed examination and interpretation. This is advantageous over traditional staining methods that require the use of chemical dyes and subsequent expert pathologist examination for evaluation.

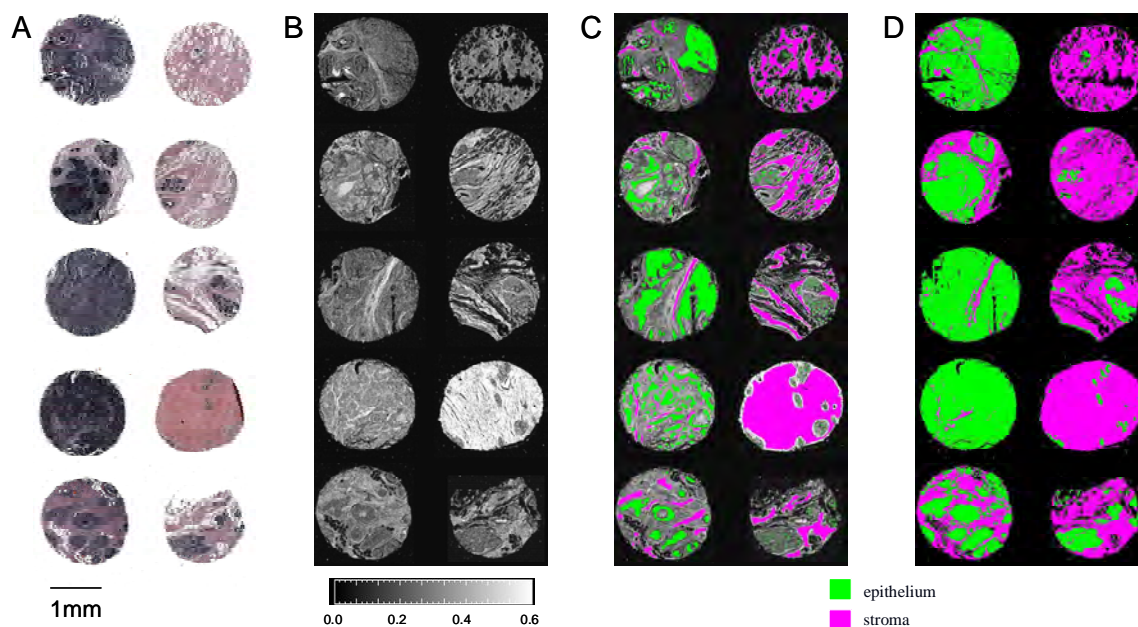


Fig. 1. Conventional H&E stained images, FT-IR spectral images and classification. (A) An H&E stained image of tissue cores from five invasive ductal carcinoma patients. Each row represents a single patient, with malignant tissue samples on the left and normal samples on the right. (B) An FT-IR image at 3292 cm^{-1} denotes the NH bending vibration at the amide A protein mode. Brighter regions denote relatively protein-rich stroma. (C) A ground truth FITR image with pixels marked as stroma or epithelium serves as the gold standard for ROC analysis and classification evaluation. (D) A classified FT-IR image in which all pixels are labeled as stroma or epithelium accurately corresponds to the H&E stained image. The classification does not require stains or human interpretation.

4. RESULTS

4.1 Effect of spectral resolution on tissue segmentation

The impact of spectral resolution on classification performance is evaluated by downsampling spectra at every pixel with a neighbor binning and interpolation procedure. FT-IR image data sets are acquired at 4 cm^{-1} spectral resolution and are downsampled to 8, 16, 32, 64, and 128 cm^{-1} resolution. As seen in Figure 2A, an average spectrum at each resolution from epithelial cells in the gold standard demonstrates that important spectral elements remain identifiable at coarser resolutions. While we anticipate that the area under the peaks would be preserved, peak shapes begin to change at a coarser spectral resolution of 32 or 64 cm^{-1} due to overlaps in the complicated spectral response. It would not be surprising to note that the most robust predictors of class incorporate best both biological diversity and spectral noise (arising from both measurement and artifacts). Hence, we anticipate that the use of these metrics would also prove robust when spectra are downsampled. Figure 2B demonstrates that the classification accuracy is not significantly affected until the spectral resolution is decreased to 128 cm^{-1} .

The result is indeed surprising as numerous prior biomedical studies with vibrational spectroscopy have employed 4 cm^{-1} to 16 cm^{-1} spectral resolution. There are two important differences between the problem here and a majority of those studies. First, many of the reported studies used sensitive spectral analysis tools (e.g. second derivatives) or were looking for fine spectral features. Second, models for pathology may have needed more complex information. Here, we are examining a 2 class problem of very distinct cell types. Hence, the acceptable classification at very coarse resolutions is likely permitted by the significant biochemical differences between stroma and epithelium in the metrics selected. Previous studies have provided evidence of clear differences in IR spectra from DNA-rich tissues such as epithelium and RNA and protein-rich tissues such as stroma,^{14,20} especially in the IR fingerprint region from $500\text{--}1500\text{ cm}^{-1}$.⁸ We hypothesize that a more complex model with additional tissue classes would likely require a higher spectral resolution for reasonable classification, but that this resolution is not required to distinguish stroma and epithelium.

A powerful feature of the algorithm we employ is the utilization of prominent spectral features for classification. Here, the features selected as classification metrics are not very sensitive to changes in spectral resolution.³⁶ Absorbance values are accurate if the peak full width at half maximum (FWHM) is not significantly less than the spectral resolution. As biological materials have broad and overlapping lineshapes, the condition holds even for very coarse resolutions. Therefore, the values of spectral metrics are not significantly altered even if some details in the spectrum are affected at coarser spectral resolutions. The center of gravity metrics used for classification are particularly robust, as they incorporate peak position and shape and are not strongly influenced by peak modifications in downsampled spectra. Care must be exercised in making this extrapolation to all data quality. For example, for poor signal to noise ratio spectra, the center of gravity calculation will be sensitive to noise.

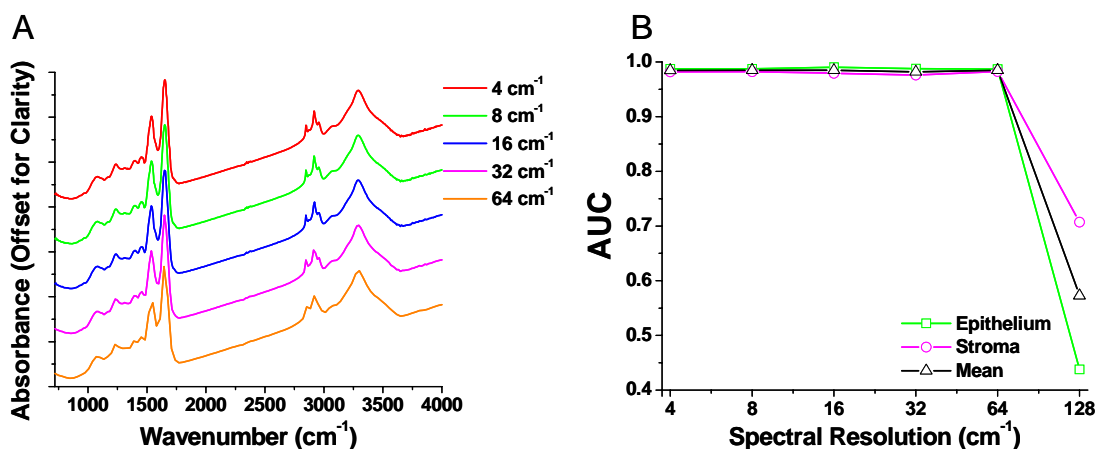


Fig. 2. Spectral resolution effect on classification. (A) Epithelial spectra obtained by downsampling data acquired at 4 cm^{-1} indicate that IR spectrum quality degrades appreciably at a spectral resolution coarser than 16 cm^{-1} , as anticipated for condensed phase biological materials. (B) AUC analysis for stroma and epithelium segmentation for each resolution demonstrates a significant decrease in classification accuracy only at a very coarse spectral resolution beyond 64 cm^{-1} .

The effective classification in downsampled FT-IR images presented in this manuscript indicates potential for faster data acquisition without significant loss in classification accuracy. Figure 2 suggests that no significant classification differences are observed in images up to 64 cm^{-1} . Since data acquisition time is estimated to decrease linearly with spectral resolution,²⁹ FT-IR images could be acquired 16 times as fast without any loss in classification performance for the two class model presented in this manuscript. Again, we emphasize that the results are preliminary and should be carefully validated. Nevertheless, the idea of optimizing data acquisition by modeling the results of other experimental conditions is an important one that should be pursued in practical translation of these protocols for clinical use.

4.2 Effect of spectral noise on tissue segmentation

Evaluation of acceptable spectral noise for FT-IR image classification is important for efficient data collection. For practical applications, it is advantageous to acquire data with the lowest SNR that permits reasonable classification. Raw data is acquired with a peak-to-peak noise value of 0.011 au, a root mean square (rms) noise value of 0.008 au, and an average amide I height of 0.328 au. To assess the impact of spectral noise on classification accuracy, Gaussian noise is added with a standard deviation of 0.001, 0.01, and 0.1 au. Figure 3 provides a qualitative evaluation of histologic images from the acquired data set (Figure 3A) and from the data sets with added Gaussian noise (Figures 3B-D).

These images indicate that acceptable classification is achieved when noise is added at a standard deviation of 0.001 au (Figure 3B), but that classification accuracy appreciably decreases with the addition of noise at or above a standard deviation of 0.01 au. This is expected, since adding noise at a standard deviation of 0.001 au does not significantly change the FT-IR image data SNR. The data set with noise added at a standard deviation of 0.01 au (Figure 3C) produces a classified image with regions of distinguishable stroma and epithelium, although there are numerous stray pixels that are not correctly classified, similar to salt and pepper noise. Upon the addition of noise of ~ 0.1 au, classified images become completely indistinguishable (Figure 3D), including the misidentification of many pixels on the empty region of the slides as tissue. This loss in classification accuracy is caused by an underlying broadening of spectral metric distributions for each class. This broadening bridges the difference in metric values. The overlap in values in turn decreases classification confidence as measured by the AUC. Hence, we have used the AUC as a reasonable measure of the classification accuracy at every experimental condition.

A plot of AUC against the added noise (Figure 3E) demonstrates that the AUC value remains relatively constant with the addition of low levels of noise. It then decreases to a mean AUC of 0.77 with the addition of noise at a standard deviation of 0.01 au and falls to a mean AUC of ~ 0.5 at a noise standard deviation of 0.1 au. It is surprising that the stroma AUC actually falls below 0.5. Though the AUC values should not be below 0.5 for classified images, our algorithm contains a pixel rejection step. A pixel is rejected if the measured metric values do not lie within the prior probability distributions. Hence, a small number of pixels are rejected at low noise levels and are not accounted.

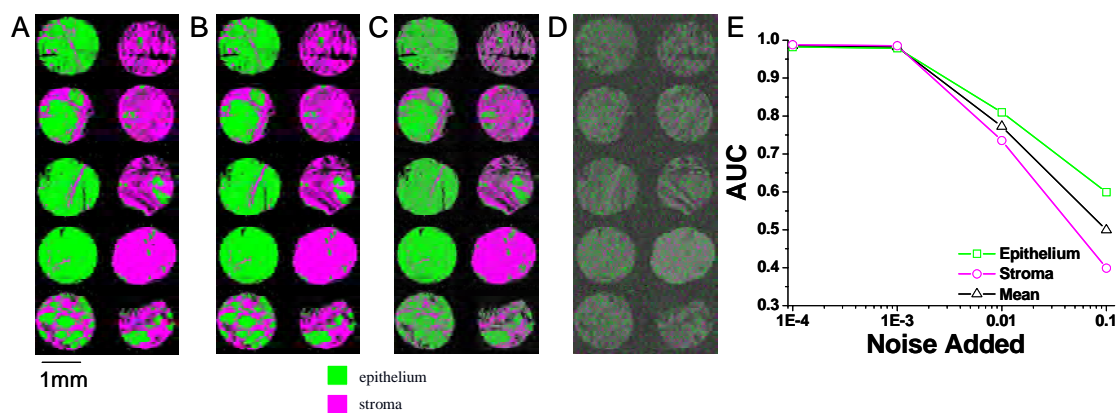


Fig. 3. Effect of noise on FT-IR image classification. Classified images are shown for (A) raw data, (B) data with Gaussian noise added at a standard deviation of 0.001 au, (C) data with Gaussian noise added at a standard deviation of 0.01 au, and (D) data with Gaussian noise added at a standard deviation of 0.1 au. (E) The AUC values for classification with noise added at a standard deviation of 0.001, 0.01, and 0.1 au confirm that classification accuracy is reasonable with a small amount of additional noise but unsatisfactory in data with a noise standard deviation at or above 0.01 au.

For the two class stroma and epithelium segmentation model presented in this manuscript, an AUC value of 0.77 does not indicate sufficient classification confidence. We would expect nearly perfect discrimination of these two types of tissue since there are numerous spectral features that distinguish epithelium and stroma.^{14,20,32,34} An estimated classification accuracy of 0.5 for this model is equivalent to random guessing and does not provide any information about tissue histology. Examination of the curve in Figure 3E indicates that some additional spectral noise at a level of 0.001 can be present without loss in classification accuracy for this two class model. We did not observe any difference in this behavior with pathology of the tissue. Breast tumor tissue is often very heterogeneous and precise pixel classification is needed to produce reasonable automated classification results. Hence these results represent a good starting point to optimize a practical protocol. There may also be a patient or clinical setting dependence of these optimal operating points that remains to be probed. From the plot, it is likely that we are close to the operating point of a practical protocol, as addition of a small amount of noise (>0.01 au) makes the classification unstable.

Last, the classification algorithm was optimized using a noise level similar to that of the acquired data set presented in this manuscript. Hence, the optimal metric sets and discriminant function are obtained for that noise level. It would not prove surprising if a de novo training and optimization of lower quality data could yield similar results. A de novo classification algorithm development, however, is not guaranteed to produce equivalent results for the higher noise cases and will fail where overlap between the prior distributions is significant due to noise broadening. Hence, we believe that the conditions found here are close to optimal.

4.3 Noise reduction with the MNF transform

In this manuscript, we have used an instrument with a high performance detector that has a low multichannel detection advantage. FT-IR imaging using large focal plane array (FPA) detectors, however, is a promising avenue for rapid data acquisitions due to the large multichannel advantage. Imaging with FPAs, unfortunately, often results in low signal-to-noise (SNR) data due to the poor detector characteristics and other limitations.³⁷ From the trading rules of FT-IR spectroscopy,²⁹ achieving a factor of n improvement in SNR would result in a increase of n^2 in data collection time. An alternative to improve SNR is to employ post-processing algorithms to reduce noise. One such avenue for noise reduction is the use of the minimum noise fraction (MNF) transform. The MNF transform can be used in a mathematical procedure to remove uncorrelated contributions from the spatial and spectral domains. First, a forward transform is used to perform a factor analysis and re-order spectral data in the order of decreasing SNR. The MNF calculation is a two-step process. A noise covariance matrix is estimated and used to decorrelate and rescale the noise in the data. Subsequently, a standard PCA performed on the noise-whitened data. A second step is to select only those factors that correspond to a sufficiently high SNR by examining the eigenvalue images. The first few eigenvalue images generally correspond to higher SNR values and contain most of the useful information. Noise reduction is achieved by suppressing the later factors corresponding largely to noise or zero-filling components and inverse transforming the data. A noise reduction by a factor greater than 5 could be achieved by this technique if the initial SNR is sufficiently high.^{38,39} Though the utility of this method is demonstrated for IR imaging,⁴⁰ its use has not been widespread. Further, the use of MNF transformed data for tissue classification has not been attempted.

We propose to use the MNF transform route as a method for fast data acquisition without loss in classification accuracy. The protocol involves rapid data collection at a low SNR, followed by application of MNF transform for noise reduction. Classification is then performed on these noise-reduced images. It must be noted that the gain here is through computational techniques and does not involve changes in instrumentation hardware or data acquisition time. A secondary advantage that may arise is that decreasing the variance in spectral data could also decrease the biologic variance in the data and should improve separation of tissue classes. Excessive image noise will broaden spectral metric distributions for each class, which increases the error associated with each metric and decreases classification confidence. Therefore, if the metric distribution mean values for each class are sufficiently different decreasing noise will decrease the area of metric distribution overlap and improve segmentation confidence.

The impact of noise reduction on classification is demonstrated in Figure 4. The MNF transform-based protocol is applied to the acquired data set and the data sets with Gaussian noise added as discussed in the previous section. Classified images are displayed for each noise level after MNF transform-aided noise reduction (Figures 4A-D). The AUC values for the MNF transformed image sets are compared with the AUC values for noisy images (Figure 4E).

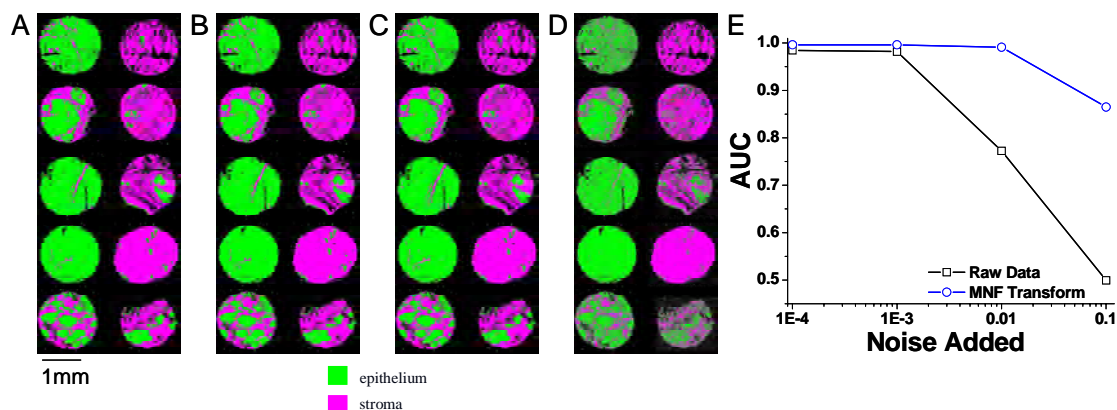


Fig. 4. Improvement in automated FT-IR image classification with the application of the MNF transform. Classified images from MNF transformed FT-IR images are shown for (A) raw data, (B) data with Gaussian noise added at a standard deviation of 0.001 au, (C) data with Gaussian noise added at a standard deviation of 0.01 au, and (D) data with Gaussian noise added at a standard deviation of 0.1 au. (E) Comparing AUC values for original FT-IR images and MNF transformed FT-IR images demonstrates that classification improves with noise reduction, especially when the noise has a standard deviation of 0.01 - 0.1 au.

Evaluation of classified images and AUC values indicates that the MNF transform improves classifier performance for each image. Given that the classification accuracy was very high, the effects of MNF transform are significant only when the noise level degrades the original data. Nevertheless, it can be seen from the figure that the high accuracy is recovered for an order of magnitude increase in data noise. Therefore, application of the MNF transform on data acquired with these noise distributions will make a significant difference in classifier performance. Specifically, we can acquire data with a noise standard deviation of 0.01 au and provide accuracy levels that are comparable to those obtained in our measurements of lower noise. This finding is significant in that noise levels of 0.01 au are commonly obtained in rapidly acquired FT-IR imaging data sets with large array detectors. Further, since the classification accuracy seems to be little affected by spectral resolution, we can anticipate that it will be little affected by the choice of an apodization function and other minor sources of error for a reasonable spectral resolution. Hence, we contend that the protocol developed here would be well-suited to rapid imaging with large array detectors.

5. CONCLUSIONS

Recent developments in FT-IR imaging and data processing facilitate new applications for this technology. In this manuscript, we report an initial application in automating histopathology of breast tissue. Supervised segmentation of breast stroma and epithelium in FT-IR images is presented and nearly-perfect classification accuracy is estimated. The impacts of spectral resolution and noise on image classification are evaluated. Results in this paper demonstrate that spectral resolution can be decreased 16-fold without loss in classification accuracy. The classification algorithm is more sensitive to noise, but noise reduction with the MNF transform can improve classification accuracy while decreasing the time required for data collection. This evaluation of the impact of experimental parameters on classification accuracy represents a first step in developing a practical protocol for rapid and automated histopathology.

REFERENCES

- [1] L. Benoit, P. Fayoulet, F. Collin, L. Arnould, J. Fraisse, and J. Cuisenier, "Histological and cytopathological cancer specimens: good practice in the operating room," *Ann. Chir.*, 128, 637-641 (2003).
- [2] A. Creager and K. Geisinger, "Intraoperative evaluation of sentinel lymph nodes for breast carcinoma: current methodologies," *Adv. Anat. Pathol.*, 9(4), 233-243 (2002).
- [3] A. Cochran, R. Huang, J. Guo, and D. Wen, "Current practice and future directions in pathology and laboratory evaluation of the sentinel node," *Ann. Surg. Oncol.*, 8(9 Suppl.), 13S-17S (2001).

- [4] M. Simunovic, A. Gagliardi, D. McCready, A. Coates, M. Levine, and D. DePetrillo, "A snapshot of waiting times for cancer surgery provided by surgeons affiliated with regional cancer centers in Ontario," *CMAJ*, 165(4), 421-425 (2001).
- [5] E. Rakovitch, A. Mihai, J. Pignol, W. Hanna, J. Kwinter, C. Chartier, I. Ackerman, J. Kim, K. Pritchard, and L. Paszat, "Is expert breast pathology assessment necessary for the management of ductal carcinoma in situ?," *Breast Cancer Res. Treat.*, 87, 265-272 (2004).
- [6] E. Newman, A. Guest, M. Helvie, M. Roubidoux, A. Chang, C. Kleer, K. Diehl, V. Cimmino, L. Pierce, D. Hayes, L. Newman, and M. Sabel, "Changes in surgical management resulting from case review at a multidisciplinary tumor board," *Cancer*, 107, 2346-2351 (2005).
- [7] D. Slamon, G. Clark, S. Wong, W. Levin, A. Ullrich, and W. McGuire, "Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene," *Science*, 235(4785), 177-182 (1987).
- [8] D. Naumann, "FT-infrared and FT-Raman spectroscopy in biomedical research," *Appl. Spec. Review*, 36(2-3), 239-298 (2001).
- [9] D. Ellis and R. Goodacre, "Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy," *Analyst*, 131, 875-885 (2006).
- [10] G. Srinivasan and R. Bhargava, "Fourier transform-infrared spectroscopic imaging: the emerging evolution from a microscopy tool to a cancer imaging modality," *Spectroscopy*, 22(7), 30-43 (2007).
- [11] I.W. Levin and R. Bhargava, "Fourier transform infrared vibrational spectroscopic imaging: integrating microscopy and molecular recognition," *Annu. Rev. Phys. Chem.*, 56, 429-474 (2005).
- [12] M. Diem, M. Romeo, S. Boydston-White, M. Miljkovic, and C. Mattheus, "A decade of vibrational micro-spectroscopy of human cells and tissue (1994-2004)," *Analyst*, 129, 880-885 (2004).
- [13] H. Fabian, N. Thi, M. Eiden, P. Lasch, J. Schmitt, and D. Naumann, "Diagnosing benign and malignant lesions in breast tissue sections by using IR-microspectroscopy," *Biochim. Biophys. Acta*, 1758(7), 874-882 (2006).
- [14] H. Fabian, M. Jackson, L. Murphy, P. Watson, I. Fichtner, and H.H. Mantsch, "A comparative infrared spectroscopic study of human breast tumors and breast tumor cell xenografts," *Biospectroscopy*, 1, 37-45 (1995).
- [15] Y. Ci, T. Gao, J. Feng and Z. Guo, "Fourier transform infrared spectroscopic characterization of human breast tissue: implications for breast cancer diagnosis," *Appl. Spec.*, 53(3), 312-315 (1999).
- [16] C. Liu, Y. Zhang, X. Yan, X. Zhang, C. Li, W. and Yang, and D. Shi, "Infrared absorption of human breast tissues in vitro," *J. Lumin.*, 119-120, 132-136 (2006).
- [17] C. Petibois, and G. Deleris, "Chemical mapping of tumor progression by FT-IR imaging: towards molecular histopathology," *Trends Biotechnol.*, 24(10), 455-462 (2006).
- [18] M. Jackson, J. Mansfield, B. Dolenko, R. Somorjai, H.H. Mantsch, and P. Watson, "Classification of breast tumors by grade and steroid receptor status using pattern recognition analysis of infrared spectra," *Cancer Detect. Prev.*, 23(3), 245-253 (1999).
- [19] R. Eckel, H. Huo, W. Guan, X. Hu, X. Che, and W.D. Huang, "Characteristic infrared spectroscopic patterns in the protein bands of human breast tissue," *Vib. Spec.*, 27, 165-173 (2001).
- [20] M. Diem, S. Boydston-White, and L. Chiriboga, "Infrared spectroscopy of cells and tissues: shining light onto a novel subject," *Appl. Spec.*, 53(4), 148A-161A (1999).
- [21] A. Haka, K. Shafer-Peltier, M. Fitzmaurice, J. Crowe, R.R. Dasari, and M.S. Feld, "Diagnosing breast cancer by using Raman spectroscopy," *Proc. Natl. Acad. Sci. USA*, 102(35), 12371-12376 (2005).
- [22] P. Matousek and N. Stone, "Prospects for the diagnosis of breast cancer by noninvasive probing of calcifications using transmission Raman spectroscopy," *J. Biomed. Opt.*, 12(2), 024008 (2007).
- [23] C. Adem, C. Reynolds, J. Ingle, and A.. Nascimento, "Primary breast sarcoma: clinicopathologic series from the Mayo Clinic and review of the literature," *Br. J. Cancer*, 91(2), 237-241 (2004).
- [24] E. Blout, and R. Mellors, "Infrared Spectra of Tissues," *Science*, 110, 137-138 (1949).
- [25] E.N. Lewis, P.J. Treado, R.C. Reeder, G. Story, A. Dowrey, C. Marcott, and I.W. Levin, "Fourier transform spectroscopic imaging using an infrared focal-plane array detector," *Anal. Chem.*, 67, 3377-3381 (1995).
- [26] J. Kononen, L. Bubendorf, A. Kallioniemi, M. Barlund, P. Schraml, S. Leighton, J. Torhorst, M. Mihatsch, G. Sauter, and O. Kallioniemi, "Tissue microarrays for high-throughput molecular profiling of tumor specimens," *Nat. Med.*, 4(7), 844-847 (1998).
- [27] D.C. Fernandez, R. Bhargava, S.M. Hewitt, and I.W. Levin, "Infrared spectroscopic imaging for histopathologic recognition," *Nat. Biotechnol.*, 23(4), 469-474 (2005).

- [28] R. Bhargava, D.C. Fernandez, S.M. Hewitt, and I.W. Levin, "High throughput assessment of cells and tissues: Bayesian classification of spectral metrics from infrared vibrational spectroscopic imaging data," *Biochim. Biophys. Acta*, 1758(7), 830-845 (2006).
- [29] P.R. Griffiths and J. De Haseth, *Fourier Transform Infrared Spectroscopy*, John Wiley & Sons, New York, 1986.
- [30] A. Green, M. Berman, P. Switzer, and M. Craig, "A transformation for ordering multispectral data in terms of image quality with implications for noise removal," *IEEE Trans. Geosci. Remote Sensing*, 26, 65-74 (1988).
- [31] P.P. Rosen, *Rosen's Breast Pathology*, Lippincott, Williams, and Wilkins, Philadelphia, 2001.
- [32] F.N. Keith and R. Bhargava, "Data processing for tissue histopathology using Fourier transform infrared spectra," *Proc. Asilomar Conference on Systems, Signals, and Computers*, 71-75 (2006).
- [33] F.N. Keith and R. Bhargava, "Towards automated breast histopathology using mid-IR spectroscopic imaging," *Technol. Cancer Res. Treat.*, in preparation.
- [34] M. Jackson, L. Choo, P. Watson, W. Halliday, and H.H. Mantsch, "Beware of connective tissue proteins: assignment and implications of collagen absorptions in infrared spectra of human tissues," *Biochim. Biophys. Acta*, 1270, 1-6 (1995).
- [35] R. Salzer, G. Steiner, H.H. Mantsch, J. Mansfield, and E.N. Lewis, "Infrared and Raman imaging of biological and biomimetic samples," *Fresenius J. Anal. Chem.*, 366, 712-726 (2000).
- [36] R. Bhargava, "Towards a practical Fourier transform infrared chemical imaging protocol for cancer histopathology," *Anal. Bioanal. Chem.*, 389, 1155-1169 (2007).
- [37] C.M. Snively and J.L. Koenig, "Characterizing the performance of a fast FT-IR imaging spectrometer," *Appl. Spec.*, 53(2), 170-177 (1999).
- [38] R. Bhargava, S.Q. Wang, and J.L. Koenig, "Route to higher fidelity FT-IR imaging," *Appl. Spec.*, 54(4), 486-495 (2000).
- [39] R. Bhargava, T. Ribar, and J.L. Koenig, "Towards faster FT-IR imaging by reducing noise," *Appl. Spec.*, 53(11), 1313-1322 (1999).
- [40] M. Wabomba, Y. Sulub, and G. Small, "Remote Detection of Volatile Organic Compounds by Passive Multispectral Infrared Imaging Measurements," *Appl. Spec.*, 61(4), 349-358 (2007).